On monotonicity, stability, and construction of central schemes for hyperbolic conservation laws with source terms (Revised Version)

V. S. Borisov * M. Mond †

November 4, 2008

Abstract

The monotonicity and stability of difference schemes for, in general, hyperbolic systems of conservation laws with source terms are studied. The basic approach is to investigate the stability and monotonicity of a non-linear scheme in terms of its corresponding scheme in variations. Such an approach leads to application of the stability theory for linear equation systems to establish stability of the corresponding non-linear scheme. The main methodological innovation is the theorems establishing the notion that a non-linear scheme is stable (and monotone) if the corresponding scheme in variations is stable (and, respectively, monotone). Criteria are developed for monotonicity and stability of difference schemes associated with the numerical analysis of systems of partial differential equations. The theorem of Friedrichs (1954) is generalized to be applicable to variational schemes with non-symmetric matrices. A new modification of the central Lax-Friedrichs (LxF) scheme is developed to be of the second order accuracy. A monotone piecewise cubic interpolation is used in the central schemes to give an accurate approximation for the model in question. The stability and monotonicity of the modified scheme are investigated. Some versions of the modified scheme are tested on several conservation laws, and the scheme is found to be accurate and robust. As applied to hyperbolic conservation laws with, in general, stiff source terms, it is constructed a second order scheme based on operator-splitting techniques.

1 Introduction

We are mainly concerned with the stability and monotonicity [10] of difference schemes for hyperbolic systems of conservation laws with source terms. Such

*The Pearlstone Center for Aeronautical Engineering Studies, Department of Mechanical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. E-mail: viatslav@bgu.ac.il

†The Pearlstone Center for Aeronautical Engineering Studies, Department of Mechanical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. E-mail: mond@bgu.ac.il

systems are used to describe many physical problems of great practical importance in magneto-hydrodynamics, kinetic theory of rarefied gases, linear and nonlinear waves, viscoelasticity, multi-phase flows and phase transitions, shallow waters, etc. (see, e.g., [7], [12], [21], [30], [36], [37], [40], [42], [46], [51], [52]). We will consider a 1-D system of the conservation laws written in the following form

$$\frac{\partial \mathbf{u}}{\partial t}+\frac{\partial}{\partial x}\mathbf{f}\left(\mathbf{u}\right)=\frac{1}{\tau}\mathbf{q}\left(\mathbf{u}\right), \quad x\in\mathbb{R},\ 0<t\leq T_{\max}; \quad \mathbf{u}\left(x,t\right)|_{t=0}=\mathbf{u}^{0}\left(x\right), \quad (1)$$

where $\mathbf{u}=\left\{u_1,u_2,\ldots,u_M\right\}^T$ is a vector-valued function from $\mathbb{R}\times[0,+\infty)$ into an open subset $\Omega_{\mathbf{u}}\subset\mathbb{R}^M$, $\mathbf{f}\left(\mathbf{u}\right)=\left\{f_1\left(\mathbf{u}\right),f_2\left(\mathbf{u}\right),\ldots,f_M\left(\mathbf{u}\right)\right\}^T$ is a smooth function (flux-function) from $\Omega_{\mathbf{u}}$ into $\mathbb{R}^M$, $\mathbf{q}\left(\mathbf{u}\right)=\left\{q_1\left(\mathbf{u}\right),q_2\left(\mathbf{u}\right),\ldots,q_M\left(\mathbf{u}\right)\right\}^T$ denotes the source term, $\tau>0$ denotes the stiffness parameter, $\mathbf{u}^0\left(x\right)$ is either of compact support or periodic. We will assume that $\tau=const$ without loss of generality. We will assume that the system (1) is hyperbolic and, hence, the Jacobian matrix of $\mathbf{f}\left(\mathbf{u}\right)$ possesses M linearly independent eigenvectors (see, e.g., [21]). In addition, it is assumed in this paper that

$$\sup_{\mathbf{u}\in\Omega_{\mathbf{u}}}\left\|\mathbf{A}\right\|_2\leq\lambda_{\max}<\infty, \quad \mathbf{A}=\frac{\partial\mathbf{f}\left(\mathbf{u}\right)}{\partial\mathbf{u}}, \quad (2)$$

and all eigenvalues, $\xi_k=\xi_k\left(\mathbf{u}\right)$, of the Jacobian matrix $\mathbf{G}$ $(=\partial\mathbf{q}\left(\mathbf{u}\right)/\partial\mathbf{u})$ have non-positive real parts, i.e.

$$R_e\xi_k\left(\mathbf{u}\right)\leq 0, \quad \forall k, \quad \forall\mathbf{u}\in\Omega_{\mathbf{u}}. \quad (3)$$

Here and in what follows $\left\|\mathbf{M}\right\|_p$ denotes the matrix norm of a matrix $\mathbf{M}$ induced by the vector norm $\left\|\mathbf{v}\right\|_p=\left(\sum_i\left|v_i\right|^p\right)^{1/p}$, and $\left\|\mathbf{M}\right\|$ denotes the matrix norm induced by a prescribed vector norm. $\mathbb{R}$ and $\mathbb{C}$ denote the fields of real and complex numbers, respectively, and $\mathbb{K}$ denotes either of these fields.

In the numerical solution of the, in general, stiff ($\tau\ll 1$) system (1), one is seeking to establish a numerical scheme that would be robust enough to eradicate spurious oscillations, i.e. a monotone scheme [10]. At the present time there are several notions of scheme monotonicity. The notion of 'monotonicity preserving scheme' originally appeared in Godunov [22]. Such a scheme transforms a monotone increasing (or decreasing) function $v\left(x\right)$ of space coordinate x at a time level t into a monotone increasing (or decreasing, respectively) function $\widehat{v}\left(x\right)$ at the next time level $t+\Delta t$. Thus, with the use of a monotonicity preserving scheme, any discontinuity in the initial monotone data can be smeared in succeeding time steps but cannot become oscillatory. Nowadays monotonicity preserving schemes are also known as, e.g., monotonicity conserving iterations (or methods) [29], monotone schemes (e.g., [1], [34], [44]), monotonicity preserving methods [40], and Godunov-monotone schemes [10]. Harten et al. [25] put forward their own definition of scheme monotonicity as follows: a difference scheme, $\widehat{v}_i=H\left(v_{i-k},v_{i-k+1},\ldots,v_{i+k}\right)$, is said to be monotone if H is a monotone increasing function of each of its arguments. The following definition is due to Samarskiy: a scheme is regarded as monotone if the boundary

maximum principle is maintained [58] (see also, e.g., [59, p. 183], [10], [8]). A difference scheme may also be referred to as monotone if a maximum principle, e.g., the boundary maximum principle, the region maximum principle, the maximum principle for inverse column entries, the maximum principle for the absolute values, etc', holds for this scheme [11]. A further important notion of difference scheme monotonicity was, in fact, done in [50] (see also [10]). A scheme will be referred to as monotone if it is monotonicity preserving [22] and transforms a "$\wedge$-function" (or "$\vee$-function") into a "μ - function" (or into an "η - function", respectively). Here and in what follows, a scalar grid function v_i will be referred to as $\wedge$-function (or $\vee$-function) if there exist grid nodes m and n such that $m \leq n$; $v_m > v_{m-1}$ and $v_i \geq v_{i-1}$ ($v_m < v_{m-1}$ and $v_i \leq v_{i-1}$ for the $\vee$-function) if $i < m$; $v_i = const$ if $m \leq i \leq n$; $v_n > v_{n+1}$ and $v_i \geq v_{i+1}$ ($v_n < v_{n+1}$ and $v_i \leq v_{i+1}$ for the $\vee$-function) if $i > n$. Simply stated, the $\wedge$-function (or $\vee$-function) is a scalar grid function v_i that has only one generalized local maximum [50] (or generalized local minimum [50], respectively). The set of μ-functions (or η-functions) is the union of $\wedge$-functions ($\vee$-functions, respectively) and the set of monotone functions.

Hereinafter, for the sake of convenience, a monotonicity preserving scheme will be referred to as G-monotone for short, a scheme monotone in terms of Harten et al. [25] will be referred to as H-monotone, a scheme will be referred to as S-monotone if a maximum principle holds for this scheme (see, e.g., [58], [59, p. 183], [10], [11]), and a scheme being monotone from the standpoint of Ostapenko [50] will be referred to as GO-monotone [10].

For studying G-monotonicity of non-linear schemes the notion of total variation (TV, see, e.g., [21], [36], [40]) turns out to be an useful tool, since any total variation diminishing (TVD) scheme is G-monotone [21, p. 168], [26], [40, p. 110]. H-monotone as well as TVD schemes appear to be attractive for at least the following reasons: (i) H-monotone schemes are TVD [26] and, hence, G-monotone [26]; (ii) any H-monotone scheme, being TVD, converges to the physically relevant solution, i.e. solutions of H-monotone schemes do satisfy the entropy condition [25], [26]; (iii) The notion of a TVD method is sufficient to prove convergence [40, p. 148]; (iv) There exist simple sufficient conditions for a scalar scheme to be TVD [21, p. 169]. Besides, it is widely believed that TVD methods are free from spurious oscillations (e.g., [13], [26], [32], [40], [41], [48]); and thus, for the above-stated reasons, TVD schemes are in common practice (see, e.g., [21], [36], [37], [40], [44], [45], [46], [51], [52], [53], [61]).

Let us note that Harten's theorem relative to G-monotonicity of H-monotone schemes (i.e., H-monotonicity $\Rightarrow$ TVD $\Rightarrow$ G-monotonicity) was proven in [26, p. 360] for a specific class of schemes approximating a 1-D scalar partial differential equation (PDE), and hence it does not always happen that an H-monotone scheme is G-monotone. Actually, let us consider the following linear scheme:

$$\widehat{v}_i = \alpha_i v_{i-1} + \beta_i v_i + \gamma_i v_{i+1}, \tag{4}$$

where $\alpha_i = 1 - 2\varepsilon$, $\beta_i = \gamma_i = \varepsilon$ at $i = 1, 2, 3, \ldots$, and $\alpha_i = \beta_i = \varepsilon$, $\gamma_i = 1 - 2\varepsilon$ at $i = 0, -1, -2, \ldots$, $(0 < \varepsilon < 0.25)$. Scheme (4) is H-monotone, since α_i, β_i,

$\gamma_i > 0\ \forall i$. Considering the monotone increasing function v_i, namely $v_i = 0$ for all $i \leq 0$ and $v_i = 1$ for all $i > 0$, we obtain that $\widehat{v}_i = 0$ for all $i < 0$, $\widehat{v}_0 = 1 - 2\varepsilon$, $\widehat{v}_1 = 2\varepsilon$, and $\widehat{v}_i = 1$ for all $i > 1$. We note that the grid function $\widehat{v}_i$ will not be monotone. Thus, H-monotonicity is not sufficient as well as not necessary (see, e.g., [10, Example 4.2]) condition for a scheme to be free of spurious oscillations.

Furthermore, as demonstrated in [10], a conservative scheme, H-monotone (hence, consistent with the entropy condition [25]), TVD (hence, G-monotone [26]), and S-monotone, be it non-linear [10, pp. 1578-1580] or even linear with constant coefficients [10, p. 1561], can produce spurious oscillations comparable with the size of the jump-discontinuity (cf. [28]). Thus, the notions of TVD, S-monotonicity, and G-monotonicity are not sufficient for a numerical scheme to be non-oscillatory. As is shown in [10], the notion of GO-monotonicity is a very helpful tool for the construction of non-oscillatory schemes. However, a GO-monotone scheme can be not TVD [10] and, hence, not S-monotone, since the notion TVD can be viewed as a concept within S-monotonicity [10]. Hence, if a numerical scheme will be GO-monotone as well as S-monotone (GOS-monotone for short), then this scheme will be stable and, in general, free of spurious oscillations [10].

An approach to investigate non-linear difference schemes for S-monotonicity in terms of corresponding variational schemes was suggested in [10], [11]. The advantage of such an approach is that the variational scheme will always be linear and, hence, enables the investigation of the monotonicity for nonlinear operators using linear patterns. It is proven for the case of explicit schemes that the S-monotonicity of a variational scheme will guarantee that its original scheme will also be S-monotone [10]. Analogous theorem for the case of implicit schemes can be found in Section 2.1, Theorem 2. Since a variational scheme carries such an important properties of its original scheme as S-monotonicity and GO-monotonicity ([10], see also Section 2.1, Theorem 2), the following definition will be useful in the investigation on scheme monotonicity.

Definition 1 *A numerical scheme is termed variationally monotone if its variational scheme is monotone.*

The notion of TV turns out to be an effective tool, [40, p. 148], for studying the stability of non-linear schemes. Actually, the following property

$$\|\mathcal{N}(\mathbf{v} + \delta\mathbf{v}) - \mathcal{N}(\mathbf{v})\| \leq (1 + \alpha\Delta t)\,\|\delta\mathbf{v}\| \tag{5}$$

is sufficient for stability of a two-step method [40], however it is, in general, difficult to obtain. Here Δt denotes the time increment, α is a constant independent of Δt as $\Delta t \to 0$, $\mathbf{v}$ and $\delta\mathbf{v}$ are any two grid functions ($\delta\mathbf{v}$ will often be referred to as the variation of the grid function $\mathbf{v}$), $\mathcal{N}$ denotes the scheme operator. At the same time, the stability of linearized version of the non-linear scheme is generally not sufficient to prove convergence [27], [40]. Instead, the TV-stability adopted in [27] (see also [40, s. 8.3.5]) makes it possible to prove convergence (to say, TV-convergence) of non-linear scalar schemes with ease. However, the TVD property is a purely scalar notion that cannot, in general,

be extended for non-linear systems of equations, as the true solution itself is usually not TVD [21], [40]. Moreover, one can see in [10, pp. 1578-1581] that a TVD scheme can be non-convergent in, at least, L_∞, in spite that the scheme is TV-stable (see, e.g., [40, Theorem 12.2]). Such a phenomenon is caused by the fact that TV is not a norm, but a semi-norm.

Nowadays, there exists a few methods for stability analysis of some classes of nonlinear difference schemes approximating systems of PDEs (see, e.g., [18], [20], [40], [44], [47], [59] and references therein). It is noted in [20] that the problem of stability analysis is still one of the most burning problems, because of the absence of its complete solution. In particular, as noted in [18] in this connection, the vast majority of difference schemes, currently in use, have still not been analyzed. LeVeque [40] noted as well that, in general, no numerical method for non-linear systems of equations has been proven to be stable. There is not even a proof that the first-order Godunov method converges on general systems of non-linear conservation laws [40, p. 340]. Thus, a different approach to testing scheme stability must be adopted to prove convergence of non-linear schemes for systems of PDEs. The notion of scheme in variations (or variational scheme [10], [11]) has, in all likelihood, much potential to be an effective tool for studying stability of nonlinear schemes. Such an approach goes back to the one suggested by Lyapunov (1892), namely, to investigate stability by the first approximation. This idea has long been exploited for investigation of the stability of motion [19]) as well as the stability of difference equations [20]. We establish the notion that the stability of a scheme in variations implies the stability of its original scheme (see Section 2.1, Theorem 3 and Remark 4).

Aiming to demonstrate potentialities such notions as 'variational scheme' and 'GOS-monotonicity' in construction numerical schemes, we will develop a novel version of the central Lax-Friedrichs (LxF) scheme (e.g., [21, p. 170]) with second-order accuracy in space and time. The stability of this scheme is proven in Section 4.2. We restrict our proof mainly to the case when the solutions to (1) are smooth. Notice, such a property as smoothness is peculiar not only to vanishing-viscosity (e.g., [21], [40]) solutions of a hyperbolic system. Since the stiffness parameter, τ, in the system (1) is like a viscosity [40], a solution to the system of conservation laws (1) is expected to be smooth provided that input data are sufficiently smooth. This specific type of systems is interesting itself, as such systems are not uncommon in practice. For a detailed discussion on this subject, including the sufficient conditions for the global existence of smooth solutions, see [24].

An extensive literature is devoted to central schemes, since these schemes are attractive for various reasons: no Riemann solvers, characteristic decompositions, complicated flux splittings, etc., must be involved in construction of a central scheme (see, e.g., [6], [37], [38], [40], [51], [53] and references therein), and hence such schemes can be implemented as a black-box solvers for general systems of conservation laws [37]. Let us, however, note that the numerical domain of dependence [40, p. 69] for a central scheme approximating, e.g., a scalar transport equation coincides with the numerical domain of dependence for a standard explicit scheme approximating diffusion equations [40, p. 67]. Such

a property is inherent to central schemes in contrast to, e.g., the first-order upwind schemes [40, p. 73]. Hence, central schemes do not satisfy the long known principle (e.g., [3, p. 304]) that derivatives must be correctly treated using type-dependent differences, and hence there is a risk for every central scheme to exhibit spurious solutions. The results of simulations in [48] can be seen as an illustration of the last assertion. Notice, all versions of the, so called, Nessyahu-Tadmor (NT) central scheme, in spite of sufficiently small CFL number ($Cr = 0.475$), exhibit spurious oscillations in contrast to the second-order upwind scheme ($Cr = 0.95$). The first order LxF scheme exhibits the excessive numerical viscosity. Thus, the central scheme should be chosen with great care to reflect the true solution and to avoid significant but spurious peculiarities in numerical solutions.

Let us note that LxF scheme – the forerunner for central schemes [6], [37] – does not produce spurious oscillations. While, from the pioneering works of Nessyahu and Tadmor [48] and on, the higher order versions of LxF scheme can produce spurious oscillations. The reason has to do with a negative numerical viscosity introduced to obtain a higher order accurate scheme. Let us illustrate it with the scheme (135) in [9], which is $O(\Delta t + (\Delta x)^2)$ accurate. We rewrite this scheme to read

$$\frac{1}{2}\frac{\mathbf{v}_{i+0.5}^{n+0.25} - \mathbf{v}_{i+0.5}^{n}}{0.25\Delta t} + \frac{\mathbf{f}\left(\mathbf{v}_{i+1}^{n}\right) - \mathbf{f}\left(\mathbf{v}_{i}^{n}\right)}{\Delta x} =$$

$$\frac{\Delta x^2}{\Delta t}\frac{\mathbf{v}_{i}^{n} - 2\mathbf{v}_{i+0.5}^{n} + \mathbf{v}_{i+1}^{n}}{\Delta x^2} - \frac{\Delta x^2}{4\Delta t}\frac{\mathbf{d}_{i+1}^{n} - \mathbf{d}_{i}^{n}}{\Delta x}, \tag{6}$$

where $\mathbf{d}_i^n$ denotes the derivative of the interpolant at $x = x_i$. Notice, the second term in the right-hand side of (6) is, in fact, the negative numerical viscosity. Without this term, Scheme (6) would be of the first order, namely LxF scheme. Let us note that there is a possibility to increase the scheme's order of accuracy, up to $O((\Delta t)^2 + (\Delta x)^2)$, by introducing into Scheme (6) an additional non-negative numerical viscosity. Such an approach is similar to the vanishing viscosity method [21], [40], and hence possesses its advantages, yet it appears to be free of the disadvantages of this method, since the additional viscosity term is not artificial. With this approach, the second order scheme is developed in Section 4.2, where sufficient conditions for stability as well as a necessary condition for S-monotonicity of the scheme are found. The developed scheme is tested on several conservation laws in Section 5.

A stable numerical scheme may yield spurious results when applied to a stiff hyperbolic system with relaxation (see, e.g., [2], [5], [7], [12], [14], [30], [54], [55]). Specifically, spurious numerical solution phenomena may occur when underresolved numerical schemes (i.e., insufficient spatial and temporal resolution) are used (e.g., [2], [30], [32], [46]). However, during a computation, the stiffness parameter may be very small, and, hence, to resolve the small stiffness parameter, we need a huge number of time and spatial increments, making the computation impractical. Hence, we are interested to solve the system, (1), with underresolved numerical schemes. It is significant that for relaxation systems a

numerical scheme must possess a discrete analogy to the continuous asymptotic limit, because any scheme violating the correct asymptotic limit leads to spurious or poor solutions (see, e.g., [12], [30], [31], [46], [51]). Most methods for solving such systems can be described as operator splitting ones, [14], or methods of fractional steps, [7]. After operator splitting, one solves the advection homogeneous system, and then the ordinary differential equations associated with the source terms. As reported in [23], this approach is well suited for the stiff systems. Let us however note that the schemes based on the operator-splitting techniques are of the first order in time, excluding rare cases such as, e.g., the Strang splitting [40]. Thus, following Pareschi [51, p. 1396], we conclude that such schemes are, in general, not robust. Fortunately, as applied to System (1), the second order schemes can be constructed on the basis of operator-splitting techniques with ease. We are mainly concerned with such an approach in Section 4.3.

2 Monotonicity and stability of difference schemes

2.1 Non-linear schemes

We consider a nonlinear implicit scheme

$$\mathbf{H}_i\left(\mathbf{y}_1,\ldots,\mathbf{y}_M\right)=\mathbf{x}_i,\quad i\in\omega\equiv\{1,2,\ldots,M\},\tag{7}$$

where $\mathbf{y}_i{\in}L\equiv\mathbb{K}^N$ denotes the sought-for vector-valued function of grid nodes, $\mathbf{x}_i{\in}L\equiv\mathbb{K}^N$ denotes the prescribed vector-valued function of grid nodes, $\mathbf{H}_i{=}\{H_{i,1},\ldots,H_{i,N}\}^T$ is a vector-valued function with the range belonging to $\mathbb{K}^N$. If we introduce the additional notation $\mathbf{y}{=}\left\{\mathbf{y}_1^T,\ldots,\mathbf{y}_M^T\right\}^T$, $\mathbf{x}{=}\left\{\mathbf{x}_1^T,\ldots,\mathbf{x}_M^T\right\}^T$, $\mathbf{H}{=}\left\{\mathbf{H}_1^T,\ldots,\mathbf{H}_M^T\right\}^T$, then the scheme (7) can be represented in the form

$$\mathbf{H}\left(\mathbf{y}\right)=\mathbf{x},\quad \mathbf{x}\in L^M,\ \mathbf{y}\in L^M.\tag{8}$$

Theorem 2 *Let a nonlinear implicit scheme (7) be written in the form (8), where* $\mathbf{x}\in\Omega_x\subset L^M$, Ω_x *denotes a closed and bounded convex set. Let the mapping* $\mathbf{H}$ *in (8) have a strong Fréchet derivative (strong F-derivative [49, item 3.2.9]),* $\mathbf{H}'\left(\mathbf{y}\right)$, *at every* $\mathbf{y}\in int\left(\Omega_y\right)$ *provided that* $\mathbf{H}\left(\Omega_y\right)=\Omega_x$, *and let* $\mathbf{H}'\left(\mathbf{y}\right)$ *be nonsingular. Then, for any* $\mathbf{x}$, $\mathbf{x}{+}\delta\mathbf{x}\in\Omega_x$ *the scheme will be S-monotone if its variational difference scheme is S-monotone.*

Proof. The scheme (8) can be seen, [11, p. 1126], as a two-node implicit scheme, and its variational scheme becomes

$$\delta\mathbf{x}=\mathbf{H}'\left(\mathbf{y}\right)\cdot\delta\mathbf{y},\quad \mathbf{H}'\left(\mathbf{y}\right)\equiv\frac{\partial\mathbf{H}\left(\mathbf{y}\right)}{\partial\mathbf{y}}.\tag{9}$$

As $\mathbf{H}'\left(\mathbf{y}\right)$ is nonsingular, we may rewrite (9) in the form

$$\delta\mathbf{y}=\left(\mathbf{H}'\right)^{-1}\left(\mathbf{y}\right)\cdot\delta\mathbf{x}.\tag{10}$$

Then, by (10)

$$\|\delta\mathbf{y}\| = \left\|\left(\mathbf{H}'\right)^{-1}(\mathbf{y})\cdot\delta\mathbf{x}\right\| \le \left\|\left(\mathbf{H}'\right)^{-1}(\mathbf{y})\right\|\ \|\delta\mathbf{x}\|. \tag{11}$$

Let (9) be S-monotone, i.e. let

$$\|\delta\mathbf{y}\| \le \|\delta\mathbf{x}\|. \tag{12}$$

In view of (11) and (12) we obtain [11, p. 1126] that

$$\left\|\left(\mathbf{H}'\right)^{-1}(\mathbf{y})\right\| \le 1,\ \forall\mathbf{y}\in\Omega_y \subseteq L^M. \tag{13}$$

In view of the inverse function theorem [49, item 5.2.1] we obtain from (8) that

$$\mathbf{y} = \mathbf{H}^{-1}(\mathbf{x}), \quad \mathbf{x}\in\Omega_x \subset L^M,\ \mathbf{y}\in\Omega_y \subseteq L^M. \tag{14}$$

By virtue of the mean-value theorem [49, item 3.2.3] we obtain from (14)

$$\|d\mathbf{y}\| = \left\|\mathbf{H}^{-1}(\mathbf{x}+d\mathbf{x}) - \mathbf{H}^{-1}(\mathbf{x})\right\| \le \sup_{0\le t\le 1}\left\|\left(\mathbf{H}^{-1}\right)'(\mathbf{x}+td\mathbf{x})\right\|\ \|d\mathbf{x}\|$$

$$\le \sup_{\mathbf{y}\in\Omega_y}\left\|\left(\mathbf{H}^{-1}\right)'(\mathbf{H}(\mathbf{y}))\right\|\ \|d\mathbf{x}\|. \tag{15}$$

In view of the inverse function theorem [49, item 5.2.1] we can write

$$\left(\mathbf{H}^{-1}\right)'(\mathbf{H}(\mathbf{y})) = \left(\mathbf{H}'\right)^{-1}(\mathbf{y}). \tag{16}$$

By virtue of (16) and (13) we obtain from (15) that

$$\|d\mathbf{y}\| \le \|d\mathbf{x}\|. \tag{17}$$

The inequality (17) manifests the prove of the theorem. ■

Theorem 3 *Let a non-linear explicit scheme be written in the form*

$$\widehat{\mathbf{v}} = \mathbf{H}(\mathbf{v}), \quad \widehat{\mathbf{v}}\in L, \quad \mathbf{v}\in\Omega\subset L, \tag{18}$$

where Ω denotes a closed and bounded convex set in a linear vector space L. Then for any $\mathbf{v}, \mathbf{v}+\delta\mathbf{v}\in\Omega$ the scheme will be stable if its variational scheme is stable.

Proof. The variational scheme corresponding to the scheme (18) reads

$$\delta\widehat{\mathbf{v}} = \mathbf{H}'(\mathbf{v})\cdot\delta\mathbf{v}, \quad \mathbf{H}'(\mathbf{v}) \equiv \frac{\partial\mathbf{H}(\mathbf{v})}{\partial\mathbf{v}}. \tag{19}$$

The linear scheme (19) will be stable [40, p. 145] if $\|\mathbf{H}'(\mathbf{v})\| \le 1+\alpha\Delta t$ for all $\mathbf{v}\in\Omega$, that is

$$\sup_{\mathbf{v}\in\Omega}\|\mathbf{H}'(\mathbf{v})\| \le 1+\alpha\Delta t. \tag{20}$$

By virtue of the mean-value theorem [49, item 3.2.3] we obtain from (18)

$$\|\mathbf{H}(\mathbf{v}+\delta\mathbf{v})-\mathbf{H}(\mathbf{v})\| \leq \sup_{0\leq\zeta\leq 1}\|\mathbf{H}'(\mathbf{v}+\zeta\delta\mathbf{v})\|\,\|\delta\mathbf{v}\| \leq \sup_{\mathbf{v}\in\Omega}\|\mathbf{H}'(\mathbf{v})\|\,\|\delta\mathbf{v}\|. \quad (21)$$

In view of (20) we conclude from (21) that the inequality (5) for (18) will be fulfilled, and hence the original non-linear scheme (18) will be stable. ■

Remark 4 *Theorem 3 can be reformulated for implicit schemes with ease. The proof of this theorem is identical to the proof of Theorem 2.*

2.2 Linear schemes

We will consider explicit linear schemes on a uniform grid with time step Δt and spatial mesh size Δx. In view of the CFL condition [40], we assume for the explicit schemes, that $\Delta t = O(\Delta x)$. Moreover, we will also assume that $\Delta x = O(\Delta t)$, since a central scheme generates a conditional approximation to Eq. (1) (see Section 3). In such a case, the following inequalities will be valid, for sufficiently small Δt and Δx,

$$\nu_0\Delta t \leq \Delta x \leq \mu_0\Delta t, \quad \nu_0,\mu_0 = const,\ 0<\nu_0\leq\mu_0. \quad (22)$$

Notice, for hyperbolic problems it is often assumed that Δt and Δx are related in a fixed manner (e.g., [40, p. 140], [57, p. 120]), i.e. it is assumed that Δt and Δx fulfill a more strong condition than (22).

Let $\mathbf{v} \equiv \{\ldots,\mathbf{v}_i^T,\ldots\}^T$ (or $\mathbf{A}=\{\mathbf{A}_{ij}\}$) be a partitioned [43] vector (or matrix, respectively), then we shall denote by $\langle\mathbf{v}\rangle$ the ordinary vector obtained from $\mathbf{v}$ (or by $\langle\mathbf{A}\rangle$ the ordinary matrix obtained from $\mathbf{A}$, respectively) by removing its partitions. It is easy to see that

$$\|\mathbf{v}\|_\infty \equiv \max_i \|\mathbf{v}_i\|_\infty = \|\langle\mathbf{v}\rangle\|_\infty. \quad (23)$$

To start with, we obtain necessary conditions for some class of linear schemes to be GOS-monotone. We consider the following explicit homogeneous scheme

$$\mathbf{z}_i = \sum_j \mathbf{B}_{ij}\cdot\mathbf{y}_j, \quad \mathbf{z}_i,\mathbf{y}_j \in L, \quad (24)$$

where L denotes the linear vector space with the orthonormal basis $\{\mathbf{b}_l\}_1^M$, $\mathbf{b}_1 = \{1,0,\ldots,0\}^T$, $\mathbf{b}_2 = \{0,1,\ldots,0\}^T$, ..., $\mathbf{b}_M = \{0,0,\ldots,1\}^T$; $\mathbf{B}_{ij} \equiv \{B_{ij}^{kl}\}$ is a square matrix. It is assumed that any constant (i.e., $\mathbf{z}_i = \mathbf{y}_j \equiv \mathbf{c} = const$) is a solution to (24). Then, in view of (24), we find that

$$\sum_j \mathbf{B}_{ij} = \mathbf{I}, \quad \forall i \quad (25)$$

will be the necessary conditions for (24) to be G-monotone (cf. [10, p. 1560]). Here and in what follows, $\mathbf{I}$ denotes the identity operator.

Theorem 5 *Let an explicit linear scheme be written in the form (24), and let any constant be a solution to (24). If (24) is GOS-monotone, then the diagonal elements, B_{ij}^{kk}, of the matrices $\mathbf{B}_{ij} \equiv \left\{B_{ij}^{kl}\right\}$ are non-negative, i.e.*

$$B_{ij}^{kk} \geq 0, \quad \forall i, j, k, \tag{26}$$

and

$$B_{ij}^{kk} \text{ is a } \mu - \text{function of } i \text{ for all } j \text{ and } k. \tag{27}$$

Proof. We consider (24) when $\mathbf{y}_j = y_j \mathbf{b}_l$, $l = 1, 2, ..., M$, where y_j is a scalar value. Let $\left\{z_{1,i}^l, z_{2,i}^l, \ldots, z_{M,i}^l\right\}^T$ be the left-hand side of (24) under $\mathbf{y}_j = y_j \mathbf{b}_l$. Then we obtain from (24) the following system of decoupled scalar equalities

$$z_{k,i}^l = \sum_j B_{ij}^{kl} y_j, \quad k, l = 1, 2, \ldots, M. \tag{28}$$

In view of Corollary 3.14 in [10], the scheme (28) will be S-monotone *iff*

$$\sum_j \left|B_{ij}^{kl}\right| \leq 1, \quad \forall i, k, l. \tag{29}$$

In view of (25) we have

$$\sum_j B_{ij}^{kk} = 1, \quad \forall i, k, \tag{30}$$

and hence

$$\sum_j \left|B_{ij}^{kk}\right| \geq 1, \quad \forall i, k. \tag{31}$$

By virtue of (31) and using (29) under $k = l$ we obtain that

$$\sum_j \left|B_{ij}^{kk}\right| = 1, \quad \forall i, k. \tag{32}$$

Thus, (30) and (32) must be valid simultaneously. It is possible *iff* all coefficients B_{ij}^{kk} comply with (26).

To prove (27) we consider (28) under $k = l$. Let m be the scheme matrix column number and δ_{mj} denote the Kronecker delta. Assuming that the scheme will be GO-monotone, it will transform $y_j = \delta_{mj}$ into a μ-function as δ_{mj} is a $\wedge$-function of j. Then we obtain from (28), under $k = l$, that $z_{k,i}^k = B_{im}^{kk}$. Hence, B_{im}^{kk} is a μ-function of i, $\forall k, m$. ■

Consider the special case of the scheme (24), namely, $\mathbf{B}_{ij}$ in (24) depends on a square matrix $\mathbf{A}_i$

$$\mathbf{B}_{ij} = \varphi_{ij}\left(\mathbf{A}_i\right), \quad \forall i, j, \tag{33}$$

where $\mathbf{A}_i$ is similar [43, p. 119] to a diagonal matrix $\mathbf{\Lambda}_i$, i.e. there exists a non-singular matrix $\mathbf{S}_i$ such that

$$\mathbf{S}_i^{-1} \cdot \mathbf{A}_i \cdot \mathbf{S}_i = \mathbf{\Lambda}_i \equiv diag\left\{\lambda_i^1, \lambda_i^2, \ldots, \lambda_i^M\right\}. \tag{34}$$

It is assumed that $\mathbf{B}_{ij}=0$ if $j\notin J_i\equiv\{j:\ i-k_L\leq j\leq i+k_R\}$, where k_L, k_R $=const\geq 0$. Notice, it is not assumed here that any constant (i.e., $\mathbf{z}_i=\mathbf{y}_j\equiv\mathbf{c}=const$) is bound to be a solution to (24). The following notation is used:

$$\mathbf{y}=\left\{\ldots,\mathbf{y}_j^T,\ldots\right\}^T,\ \overline{\mathbf{y}}_j=\mathbf{S}_j^{-1}\cdot\mathbf{y}_j,\ \overline{\mathbf{y}}=\left\{\ldots,\overline{\mathbf{y}}_j^T,\ldots\right\}^T,\ \overline{\mathbf{y}}_{i,j}=\mathbf{S}_i^{-1}\cdot\mathbf{y}_j,$$

$$\widetilde{\mathbf{y}}_i=\left\{\ldots,\overline{\mathbf{y}}_{i,i-k_L-1}^T,\overline{\mathbf{y}}_{i,i-k_L}^T,\ldots,\overline{\mathbf{y}}_{i,i+k_R}^T,\overline{\mathbf{y}}_{i,i+k_R+1}^T,\ldots\right\}^T,$$

$$\mathbf{B}=\{\mathbf{B}_{ij}\},\quad \overline{\mathbf{B}}_{ij}=\mathbf{S}_i^{-1}\cdot\mathbf{B}_{ij}\cdot\mathbf{S}_i,\quad \overline{\mathbf{B}}_i=\left\{\ldots,\overline{\mathbf{B}}_{ij-1},\overline{\mathbf{B}}_{ij},\overline{\mathbf{B}}_{ij+1},\ldots\right\}. \tag{35}$$

The stability for (24) provided (33)-(34) can be addressed by the following.

Lemma 6 *Consider an explicit scheme that can be written in the form (24) under (33), (34). Let $s_i=s(\mathbf{A}_i)$ be the spectrum of $\mathbf{A}_i$, and $\varphi_{ij}(\lambda)$ be represented by an absolutely convergent power series at each point $\lambda\in s_i$. Let $\mathbf{B}_{ij}=0$ in (24) if $j\notin J_i=\{j:\ i-k_L\leq j\leq i+k_R\}$. Then the scheme will be stable if*

$$\max_{\lambda\in s_i}\sum_j\left|\varphi_{ij}(\lambda)\right|\leq 1,\quad \forall i, \tag{36}$$

$$\left\|\left(\mathbf{S}_i^{-1}-\mathbf{S}_j^{-1}\right)\cdot\mathbf{S}_j\right\|_\infty\leq\Theta=const,\quad \forall i,\ \forall j\in J_i. \tag{37}$$

Proof. It is easy to see that

$$\mathbf{S}_i^{-1}\equiv\left\{\mathbf{I}+\left(\mathbf{S}_i^{-1}-\mathbf{S}_j^{-1}\right)\cdot\mathbf{S}_j\right\}\cdot\mathbf{S}_j^{-1}. \tag{38}$$

Then, in view of (37), we find $\forall i$, $\forall j\in J_i$

$$\left\|\mathbf{S}_i^{-1}\cdot\mathbf{y}_j\right\|_\infty\leq\left\|\mathbf{I}+\left(\mathbf{S}_i^{-1}-\mathbf{S}_j^{-1}\right)\cdot\mathbf{S}_j\right\|_\infty\left\|\mathbf{S}_j^{-1}\cdot\mathbf{y}_j\right\|_\infty\leq$$

$$\left\{1+\left\|\left(\mathbf{S}_i^{-1}-\mathbf{S}_j^{-1}\right)\cdot\mathbf{S}_j\right\|_\infty\right\}\left\|\mathbf{S}_j^{-1}\cdot\mathbf{y}_j\right\|_\infty\leq(1+\Theta)\left\|\mathbf{S}_j^{-1}\cdot\mathbf{y}_j\right\|_\infty. \tag{39}$$

By virtue of (35), we rewrite (24) to read

$$\overline{\mathbf{z}}_i\equiv\mathbf{S}_i^{-1}\cdot\mathbf{z}_i=\sum_j\overline{\mathbf{B}}_{ij}\cdot\left(\mathbf{S}_i^{-1}\cdot\mathbf{y}_j\right)\equiv\overline{\mathbf{B}}_i\cdot\widetilde{\mathbf{y}}_i\equiv\left\langle\overline{\mathbf{B}}_i\right\rangle\cdot\left\langle\widetilde{\mathbf{y}}_i\right\rangle,\quad \forall i, \tag{40}$$

where $\left\langle\overline{\mathbf{B}}_i\right\rangle$ and $\left\langle\widetilde{\mathbf{y}}_i\right\rangle$ denote the ordinary matrix and vector obtained from $\overline{\mathbf{B}}_i$ and $\widetilde{\mathbf{y}}_i$, respectively, by removing the partitions. Notice, $\overline{\mathbf{B}}_{ij}=0$ if $j\notin J_i$, since $\mathbf{B}_{ij}=0$ for these j. Thus, it can be written that $\overline{\mathbf{B}}_{ij}=\mathbf{S}_i^{-1}\cdot\mathbf{B}_{ij}\cdot\mathbf{S}_j$ if $j\notin J_i$, and hence $\overline{\mathbf{y}}_{i,j}=\overline{\mathbf{y}}_{j,j}=\mathbf{S}_j^{-1}\cdot\mathbf{y}_j=\overline{\mathbf{y}}_j$ $(\forall j\notin J_i)$. In view of (40) we obtain that

$$\left\|\overline{\mathbf{z}}_i\right\|_\infty\equiv\left\|\mathbf{S}_i^{-1}\cdot\mathbf{z}_i\right\|_\infty\leq\left\|\left\langle\overline{\mathbf{B}}_i\right\rangle\right\|_\infty\left\|\left\langle\widetilde{\mathbf{y}}_i\right\rangle\right\|_\infty,\quad \forall i, \tag{41}$$

The norm $\left\|\left\langle\widetilde{\mathbf{y}}_i\right\rangle\right\|_\infty$ in (41) can be estimated by virtue of (23) and (39):

$$\left\|\left\langle\widetilde{\mathbf{y}}_i\right\rangle\right\|_\infty=\left\|\widetilde{\mathbf{y}}_i\right\|_\infty\leq(1+\Theta)\max_j\left\|\mathbf{S}_j^{-1}\cdot\mathbf{y}_j\right\|_\infty=(1+\Theta)\left\|\overline{\mathbf{y}}\right\|_\infty,\ \forall i. \tag{42}$$

Let us estimate $\left\|\langle\overline{\mathbf{B}}_i\rangle\right\|_\infty$ in (41). In view of (34) $\mathbf{\Lambda}_i = \mathbf{S}_i^{-1}\cdot\mathbf{A}_i\cdot\mathbf{S}_i$. It can be verified, by induction with respect to n, that $(\mathbf{\Lambda}_i)^n = \mathbf{S}_i^{-1}\cdot(\mathbf{A}_i)^n\cdot\mathbf{S}_i$. Then, in view of Theorem 11.2.2 and Theorem 11.2.4 in [43], we find

$$\overline{\mathbf{B}}_{ij} \equiv \mathbf{S}_i^{-1}\cdot\mathbf{B}_{ij}\cdot\mathbf{S}_i = \varphi_{ij}\left(\mathbf{S}_i^{-1}\cdot\mathbf{A}_i\cdot\mathbf{S}_i\right) = \varphi_{ij}\left(\mathbf{\Lambda}_i\right). \tag{43}$$

Thus, $\overline{\mathbf{B}}_{ij}$ can be written in the form

$$\overline{\mathbf{B}}_{ij} = diag\left\{\Lambda_{ij}^1, \Lambda_{ij}^2, \ldots, \Lambda_{ij}^M\right\}, \quad \Lambda_{ij}^k = \varphi_{ij}\left(\lambda_j^k\right), \quad k = 1, 2, \ldots, M. \tag{44}$$

In view of (44), we find that

$$\left\|\langle\overline{\mathbf{B}}_i\rangle\right\|_\infty = \max_k \sum_j \left|\Lambda_{ij}^k\right| = \max_{\lambda\in s_i}\sum_j\left|\varphi_{ij}\left(\lambda\right)\right|, \quad \forall i. \tag{45}$$

By virtue of (42), (45), and (36) we obtain from (41) that

$$\left\|\mathbf{S}_i^{-1}\cdot\mathbf{z}_i\right\|_\infty \le (1+\Theta)\left\|\overline{\mathbf{y}}\right\|_\infty, \quad \forall i. \tag{46}$$

Since

$$\left\|\overline{\mathbf{z}}\right\|_\infty \equiv \max_i\left\|\mathbf{S}_i^{-1}\cdot\mathbf{z}_i\right\|_\infty, \tag{47}$$

we obtain, in view of (46), (47), that

$$\left\|\mathbf{z}\right\|_* \equiv \left\|\overline{\mathbf{z}}\right\|_\infty \le (1+\Theta)\left\|\overline{\mathbf{y}}\right\|_\infty \equiv (1+\Theta)\left\|\mathbf{y}\right\|_*. \tag{48}$$

The last inequality establishes Lemma 6. ■

We consider the following explicit linear scheme

$$\mathbf{v}_i^{n+1} = \sum_j \mathbf{B}_{ij}^n\cdot\mathbf{v}_j^n, \quad n \ge 0, \tag{49}$$

where

$$\mathbf{B}_{ij}^n = \begin{cases} \varphi_{ij}^n\left(\mathbf{A}_j^n\right), & j\in J_i \\ 0, & j\notin J_i \end{cases}, \quad \forall i,j,n, \tag{50}$$

$$J_i = \left\{j:\ i-k_L\le j\le i+k_R\right\}, \quad k_L, k_R = const, \quad \forall i,n, \tag{51}$$

k_L, k_R, denote the non-negative integer constants being independent of t, x, Δx, and Δt. It is assumed that the matrix-valued function $\mathbf{A} = \mathbf{A}(x,t)$ is Lipschitz-continuous and $\mathbf{A}$ is diagonizable, i.e. for $\mathbf{A}_i^n = \mathbf{A}(x_i, t_n)$ there exists a non-singular matrix $\mathbf{S}_i^n = \mathbf{S}(x_i, t_n)$ such that

$$\left(\mathbf{S}_i^n\right)^{-1}\cdot\mathbf{A}_i^n\cdot\mathbf{S}_i^n = \mathbf{\Lambda}_i^n \equiv diag\left\{\lambda_i^{n,1}, \lambda_i^{n,2}, \ldots, \lambda_i^{n,M}\right\}, \quad \forall i,n. \tag{52}$$

Let us note that even if $\mathbf{B}_{ij}^n = \varphi_{ij}^n\left(\mathbf{A}_i^n\right)$ in (50) and Lemma 6 be valid for the linear scheme (49) with $\Theta = O(\Delta t)$ at every time step, the scheme (49) will be "locally stable" only, i.e. any growth in error is, at most, order $O(\Delta t)$ in one time step. However, we cannot, in general, show on the basis of (48) that

$$\left\|\mathbf{v}^{N_T}\right\|_{**} \le C_T\left\|\mathbf{v}^0\right\|_*, \quad C_T = const, \tag{53}$$

where $\|\cdot\|_{**}$ and $\|\cdot\|_{*}$ denote some norms, $\mathbf{v}^n = \left\{\ldots,(\mathbf{v}_i^n)^T,\ldots\right\}^T$, N_T denotes the time level corresponding to time $T = N_T\Delta t$ over which we wish to compute. The reason is that the vector norm in (48) depends on the time level t_n, and hence we maynot apply (48) recursively to obtain (53). The stability of the system (49), can be addressed by the following.

Theorem 7 *Consider an explicit scheme that can be written in the form (49) under (50)-(52), where the functions $\varphi_{ij}^n(\mathbf{A})$ and $\mathbf{A}(x,t)$ are both Lipschitz-continuous. Let there exist $\Delta x_0 > 0$ such that the function $\varphi_{ij}^n(\lambda)$ in (50) can be represented by an absolutely convergent power series at each point of the spectrum $s_i^n = s(\mathbf{A}_i^n)$ $\forall i,n$, $\forall j \in J_i$, $\forall \Delta x \le \Delta x_0$, and let the matrix-valued functions $\mathbf{S}(x,t)$, and $\mathbf{S}^{-1}(x,t)$ in (52) can be taken such that the matrix-valued functions $\left[(\mathbf{S}_i^n)^{-1} - (\mathbf{S}^n)^{-1}(x)\right]\cdot\mathbf{S}^n(x)$ and $\left[(\mathbf{S}_i)^{-1}(t) - (\mathbf{S}_i^n)^{-1}\right]\cdot\mathbf{S}_i^n$ will be Lipschitz-continuous in space and, respectively, time $\forall i,n$. Let*

$$\left\|\left(\mathbf{S}_j^n\right)^{-1}\right\|_\infty \le \beta_{-1} = const, \quad \left\|\mathbf{S}_j^n\right\|_\infty \le \beta_0 = const, \quad \forall j,n. \tag{54}$$

Then the scheme (49) will be stable, i.e. (53) will be valid, if

$$\max_{\lambda\in s_i^n}\sum_j \left|\varphi_{ij}^n(\lambda)\right| \le 1, \quad \forall i,n. \tag{55}$$

Proof. Let $\check{\mathbf{B}}_{ij}^n = \varphi_{ij}^n(\mathbf{A}_i^n)$, and let us rewrite (49) to read

$$\mathbf{v}_i^{n+1} = \check{\mathbf{v}}_i^n + \hat{\mathbf{v}}_i^n, \quad \forall i,n, \tag{56}$$

where

$$\check{\mathbf{v}}_i^n = \sum_j \check{\mathbf{B}}_{ij}^n\cdot\mathbf{v}_j^n, \quad \forall i,n,\ \forall j\in J_i, \tag{57}$$

$$\hat{\mathbf{v}}_i^n = \sum_j \left(\mathbf{B}_{ij}^n - \check{\mathbf{B}}_{ij}^n\right)\cdot\mathbf{v}_j^n, \quad \forall i,n,\ \forall j\in J_i. \tag{58}$$

First, let us estimate the norm, $h_n(\cdot)$, of $\check{\mathbf{v}}^n$:

$$h_n(\check{\mathbf{v}}^n) \equiv \left\|\overline{\check{\mathbf{v}}}^n\right\|_\infty \equiv \max_i \left\|(\mathbf{S}_i^n)^{-1}\cdot\check{\mathbf{v}}_i^n\right\|_\infty. \tag{59}$$

Since $\left[(\mathbf{S}_i^n)^{-1} - (\mathbf{S}^n)^{-1}(x)\right]\cdot\mathbf{S}^n(x)$ and $\left[(\mathbf{S}_i)^{-1}(t) - (\mathbf{S}_i^n)^{-1}\right]\cdot\mathbf{S}_i^n$ are Lipschitz-continuous in space and time, respectively, we may write

$$\left\|\left[(\mathbf{S}_i^n)^{-1} - \left(\mathbf{S}_{i+1}^n\right)^{-1}\right]\cdot\mathbf{S}_{i+1}^n\right\|_\infty \le \beta_1\Delta x,\ \boldsymbol{\beta}_1{=}const, \quad \forall i,n, \tag{60}$$

$$\left\|\left[\left(\mathbf{S}_i^{n+1}\right)^{-1} - (\mathbf{S}_i^n)^{-1}\right]\cdot\mathbf{S}_i^n\right\|_\infty \le \beta_2\Delta t,\ \beta_2 = const, \quad \forall i,n. \tag{61}$$

By virtue of (60), we find

$$\left\|\left[(\mathbf{S}_i^n)^{-1} - \left(\mathbf{S}_j^n\right)^{-1}\right]\cdot\mathbf{S}_j^n\right\|_\infty \le \beta_3\Delta x,\ \beta_3 = const, \quad \forall i,n,\ \forall j\in J_i, \tag{62}$$

where $\beta_3 = \beta_1 \max(k_L, k_R)$. We assume for the explicit scheme (49), that $\Delta x = O(\Delta t)$ (i.e. $\exists\ \Delta t_0 > 0$, $\exists\ \alpha_0 > 0$ such that $\Delta x \le \alpha_0 \Delta t\ \forall \Delta t \le \Delta t_0$). Then we find by virtue of (62) that

$$\left\| \left\{ (\mathbf{S}_i^n)^{-1} - \left(\mathbf{S}_j^n\right)^{-1} \right\} \cdot \mathbf{S}_j^n \right\|_\infty \le \beta_4 \Delta t,\ \beta_4 = \alpha_0 \beta_3,\ \forall i, n,\ \forall j \in J_i. \tag{63}$$

The inequality (63) coincides with the assumption (37) in Lemma 6 under $\Theta = \beta_4 \Delta t$. Then, in view of Lemma 6, we obtain for the scheme (57), that

$$h_n(\check{\mathbf{v}}^n) \equiv \left\| \overline{\check{\mathbf{v}}}^n \right\|_\infty \equiv \max_i \left\| (\mathbf{S}_i^n)^{-1} \cdot \check{\mathbf{v}}_i^n \right\|_\infty \le$$

$$[1 + \beta_4 \Delta t] \max_i \left\| (\mathbf{S}_i^n)^{-1} \cdot \mathbf{v}_i^n \right\|_\infty \equiv [1 + \beta_4 \Delta t]\, h_n(\mathbf{v}^n). \tag{64}$$

Let us now estimate the norm $h_n(\hat{\mathbf{v}}^n)$. Since $\varphi_{ij}^n(\mathbf{A})$, $\mathbf{A}(x,t)$ are both Lipschitz continuous, we may write

$$\left\| \varphi_{ij}^n\left(\mathbf{A}_j^n\right) - \varphi_{ij}^n(\mathbf{A}_i^n) \right\|_\infty \le \alpha_1 \left\| \mathbf{A}_i^n - \mathbf{A}_j^n \right\|_\infty, \quad \forall i, n,\ \forall j \in J_i, \tag{65}$$

$$\left\| \mathbf{A}_i^n - \mathbf{A}_j^n \right\|_\infty \le \alpha_2 |x_j - x_i| \le \alpha_3 \Delta x, \quad \forall i, n,\ \forall j \in J_i, \tag{66}$$

where $\alpha_3 = \alpha_2 \max(k_L, k_R)$, $\alpha_1,\ \alpha_2 = const$. By virtue of (65), (66), and assuming that $\Delta x = O(\Delta t)$, we obtain

$$\left\| \mathbf{B}_{ij}^n - \check{\mathbf{B}}_{ij}^n \right\|_\infty \equiv \left\| \varphi_{ij}^n\left(\mathbf{A}_j^n\right) - \varphi_{ij}^n(\mathbf{A}_i^n) \right\|_\infty \le \alpha_4 \Delta t, \quad \forall i, n,\ \forall j \in J_i, \tag{67}$$

where $\alpha_4 = \alpha_0 \alpha_1 \alpha_3 = const$. We obtain from (58) that

$$(\mathbf{S}_i^n)^{-1} \cdot \hat{\mathbf{v}}_i^n = \sum_j (\mathbf{S}_i^n)^{-1} \cdot \left(\mathbf{B}_{ij}^n - \check{\mathbf{B}}_{ij}^n\right) \cdot \mathbf{S}_j^n \cdot \left(\mathbf{S}_j^n\right)^{-1} \cdot \mathbf{v}_j^n. \tag{68}$$

Whence, by virtue of (67) and (54), we obtain

$$\left\| (\mathbf{S}_i^n)^{-1} \cdot \hat{\mathbf{v}}_i^n \right\|_\infty \le \alpha_5 \Delta t \max_j \left\| \left(\mathbf{S}_j^n\right)^{-1} \cdot \mathbf{v}_j^n \right\|_\infty \equiv \alpha_5 \Delta t h_n(\mathbf{v}^n), \quad \forall i, n, \tag{69}$$

where $\alpha_5 = \beta_{-1} \beta_0 \alpha_4 \max(k_L, k_R) = const$. By virtue of (56), (64), and (69), we obtain

$$h_n\left(\mathbf{v}^{n+1}\right) \le [1 + \beta \Delta t]\, h_n(\mathbf{v}^n),\ \beta = \beta_4 + \alpha_5 = const, \quad \forall n. \tag{70}$$

It is easy to see that

$$\left(\mathbf{S}_i^{n+1}\right)^{-1} \equiv \left\{ \mathbf{I} + \left(\left(\mathbf{S}_i^{n+1}\right)^{-1} - (\mathbf{S}_i^n)^{-1} \right) \cdot \mathbf{S}_i^n \right\} \cdot (\mathbf{S}_i^n)^{-1}, \tag{71}$$

whence, by virtue of (61), we find

$$h_{n+1}\left(\mathbf{v}^{n+1}\right) = \max_i \left\| \left(\mathbf{S}_i^{n+1}\right)^{-1} \cdot \mathbf{v}_i^{n+1} \right\|_\infty \le$$

$$\max_i \left\| \mathbf{I} + \left(\left(\mathbf{S}_i^{n+1} \right)^{-1} - \left(\mathbf{S}_i^n \right)^{-1} \right) \cdot \mathbf{S}_i^n \right\|_\infty \left\| \left(\mathbf{S}_i^n \right)^{-1} \cdot \mathbf{v}_i^{n+1} \right\|_\infty \leq$$

$$(1 + \beta_2 \Delta t) \max_i \left\| \left(\mathbf{S}_i^n \right)^{-1} \cdot \mathbf{v}_i^{n+1} \right\|_\infty = (1 + \beta_2 \Delta t)\, h_n \left(\mathbf{v}^{n+1} \right). \tag{72}$$

In view of (70) and (72), we find

$$h_{n+1} \left(\mathbf{v}^{n+1} \right) \leq \left[1 + \gamma \Delta t \right]^2 h_n \left(\mathbf{v}^n \right), \ \gamma = \max \left(\alpha, \beta_2 \right), \quad \forall n. \tag{73}$$

Applying (73) recursively gives

$$h_{N_T} \left(\mathbf{v}^{N_T} \right) \leq \left(1 + \gamma \Delta t \right)^{2N_T} h_0 \left(\mathbf{v}^0 \right) \leq C_T h_0 \left(\mathbf{v}^0 \right), \quad C_T = \exp \left(2\gamma T \right). \tag{74}$$

The inequalities in (74) prove the theorem. ■

Let us consider the case when the operator $\mathbf{B}_{ij}^n$ in (49) depends on a matrix $\mathbf{A}_j^n$ belonging to a set of pairwise commutative diagonizable matrices:

$$\mathbf{B}_{ij}^n = \varphi_{ij}^n \left(\mathbf{A}_j^n \right), \ \mathbf{A}_j^n \cdot \mathbf{A}_k^m = \mathbf{A}_k^m \cdot \mathbf{A}_j^n, \quad \forall i, j, n, k, m. \tag{75}$$

In such a case, the S-monotonicity of the system (49), can be addressed by the following.

Theorem 8 *Consider an explicit scheme that can be written in the form (49) provided (75). Let $\varphi_{ij}^n (\lambda)$ in (75) can be represented by an absolutely convergent power series at each point of the spectrum $s_j^n = s \left(\mathbf{A}_j^n \right)$ $\forall i, j, n$. Then the scheme (49) will be S-monotone* iff

$$\max_{\lambda \in s_j^n} \sum_i \left| \varphi_{ij}^n (\lambda) \right| \leq 1, \quad \forall j, n. \tag{76}$$

Proof. As $\mathbf{A}_j^n$ belongs to the set of pair-wise permutable diagonizable matrices, the matrices of the set are simultaneously similar to diagonal matrices [43, p. 318], i.e., there exists a non-singular matrix $\mathbf{S}$ such that

$$\mathbf{S}^{-1} \cdot \mathbf{A}_j^n \cdot \mathbf{S} = \mathbf{\Lambda}_j^n \equiv diag \left\{ \lambda_j^{n,1}, \lambda_j^{n,2}, \ldots, \lambda_j^{n,M} \right\}, \quad \forall j, n. \tag{77}$$

where $\lambda_j^{n,m}$ denotes the m-th eigenvalue of $\mathbf{A}_j^n$. The following notation is used:

$$\overline{\mathbf{v}}_j^n = \mathbf{S}^{-1} \cdot \mathbf{v}_j^n, \ \overline{\mathbf{B}}_{ij}^n = \mathbf{S}^{-1} \cdot \mathbf{B}_{ij}^n \cdot \mathbf{S}, \ \overline{\mathbf{B}}^n = \left\{ \overline{\mathbf{B}}_{ij}^n \right\}, \ \mathbf{B}^n = \left\{ \mathbf{B}_{ij}^n \right\}. \tag{78}$$

By virtue of (78), we rewrite (49) to read

$$\overline{\mathbf{v}}_i^{n+1} \equiv \mathbf{S}^{-1} \cdot \mathbf{v}_i^{n+1} = \sum_j \overline{\mathbf{B}}_{ij}^n \cdot \overline{\mathbf{v}}_j^n. \tag{79}$$

Using $\overline{\mathbf{v}}^n \equiv \left\{ \ldots, \left(\overline{\mathbf{v}}_j^n \right)^T, \ldots \right\}^T$, we rewrite (79) to read

$$\overline{\mathbf{v}}^{n+1} = \overline{\mathbf{B}}^n \cdot \overline{\mathbf{v}}^n, \quad \overline{\mathbf{v}}^{n+1} \equiv \left\{ \ldots, \left(\overline{\mathbf{v}}_i^{n+1} \right)^T, \ldots \right\}^T. \tag{80}$$

In view of (80) we obtain that

$$h\left(\mathbf{v}^{n+1}\right) \equiv \left\|\langle\overline{\mathbf{v}}^{n+1}\rangle\right\|_1 \leq \left\|\left\langle\overline{\mathbf{B}}^{n}\right\rangle\right\|_1 \left\|\langle\overline{\mathbf{v}}^{n}\rangle\right\|_1 \equiv h\left(\mathbf{v}^{n}\right), \tag{81}$$

where $\left\langle\overline{\mathbf{B}}^{n}\right\rangle$ and $\langle\overline{\mathbf{v}}^{n}\rangle$ denote the ordinary matrix and vector obtained from $\overline{\mathbf{B}}^{n}$ and $\overline{\mathbf{v}}^{n}$, respectively, by removing the partitions. Let us estimate the norm of $\left\langle\overline{\mathbf{B}}^{n}\right\rangle$ in (81). Since $\varphi_{ij}^{n}(\lambda)$ can be represented by an absolutely convergent power series at each point $\lambda \in s_j^n = s\left(\mathbf{A}_j^n\right)$, we find, in view of Theorem 11.2.2 and Theorem 11.2.4 in [43], that

$$\overline{\mathbf{B}}_{ij}^{n} \equiv \mathbf{S}^{-1} \cdot \mathbf{B}_{ij}^{n} \cdot \mathbf{S} = \varphi_{ij}^{n}\left(\mathbf{S}^{-1} \cdot \mathbf{A}_j^n \cdot \mathbf{S}\right) = \varphi_{ij}^{n}\left(\mathbf{\Lambda}_j^n\right). \tag{82}$$

Thus, $\overline{\mathbf{B}}_{ij}^{n}$ can be written in the form

$$\overline{\mathbf{B}}_{ij} = diag\left\{\Lambda_{ij}^{n,1}, \Lambda_{ij}^{n,2}, \ldots, \Lambda_{ij}^{n,M}\right\},\ \Lambda_{ij}^{n,k} = \varphi_{ij}^{n}\left(\lambda_j^{n,k}\right),\ k = 1, 2, \ldots, M. \tag{83}$$

In view of (83), we obtain that

$$\left\|\left\langle\overline{\mathbf{B}}^{n}\right\rangle\right\|_1 = \max_j \left(\max_{k=1,\ldots,M} \sum_i \left|\Lambda_{ij}^{n,k}\right|\right) = \max_j \left(\max_{\lambda \in s_j^n} \sum_i \left|\varphi_{ij}^{n}(\lambda)\right|\right). \tag{84}$$

Whence, in view of (76), we find

$$\left\|\left\langle\overline{\mathbf{B}}^{n}\right\rangle\right\|_1 \leq 1, \quad \forall n. \tag{85}$$

The vector norm $h(\cdot)$ in (81) does not depend on time level. Then, in view of Proposition 3.2 in [11], the inequality (85) proves Theorem 8 ■

Proposition 9 *If (24) is a variational scheme, then (25) is, in general, not valid. Notice, Lemma 6 and Theorem 7 are proven without assumption (25). However, in addition to the Lipschitz-continuity of* $\mathbf{A}(x,t)$ *(see (33) and (50)), it is assumed in Lemma 6 and Theorem 7 that some functions of* $\mathbf{S}(x,t)$ *(see (34), (52)) are also Lipschitz-continuous. Let us note that the stability of a linear scheme can often be proven without assumption (25) as well as without additional assumptions on the continuity. To demonstrate it, let us generalize the theorem of Friedrichs (1954) (see, e.g., [57, p. 120], [60, p. 374]) to be applicable to variational schemes. We consider the following difference scheme*

$$\mathbf{y}^{n+1}(x) = \sum_{k=-m}^{m} \mathbf{B}_k^n(x) \cdot \mathbf{y}^n(x + k\Delta x),\ \mathbf{y}^n \in \mathbb{R}^M,\ \mathbf{B}_k^n \in \mathbb{R}^{M\times M}, \tag{86}$$

where $x \in (-\infty, \infty)$, $\mathbf{B}_k^n$ *is a symmetric and non-negative matrix. It is assumed that* $\mathbf{y}^n(x)$ *and* $\mathbf{B}_k^n(x)$ *are periodic (with the period equal to 1) functions of* x. *Let*

$$\mathbf{y}_j^n \equiv \mathbf{y}^n(x + j\Delta x),\ \mathbf{B}_{k,j}^n \equiv \mathbf{B}_k^n(x + j\Delta x), \tag{87}$$

and let

$$(\mathbf{u},\mathbf{v})\equiv\int\limits_0^1\left[\mathbf{u}^T(x)\cdot\mathbf{v}(x)\right]dx,\quad \|\mathbf{u}\|\equiv\sqrt{(\mathbf{u},\mathbf{u})}. \tag{88}$$

If there exist c_1, $c_2=const$ *such that*

$$\left\|\sum_k\mathbf{B}_k^n\right\|\le 1+c_1\Delta x,\quad \left\|\mathbf{B}_{k,k}^n-\mathbf{B}_k^n\right\|\le\frac{c_2}{2m+1}\Delta x, \tag{89}$$

then the scheme is stable. Notice, it is not assumed that $\sum_k\mathbf{B}_k^n(x)=\mathbf{I}$.

The proof is very little different from the proof when $\sum_k\mathbf{B}_k^n(x)=\mathbf{I}$. *Actually, in view of (86) and the first inequality in (89), we obtain*

$$\left\|\mathbf{y}^{n+1}\right\|^2=\sum_k\left(\mathbf{B}_k^n\cdot\mathbf{y}_k^n,\mathbf{y}^{n+1}\right)\le 0.5\sum_k\left[(\mathbf{B}_k^n\cdot\mathbf{y}_k^n,\mathbf{y}_k^n)+\left(\mathbf{B}_k^n\cdot\mathbf{y}^{n+1},\mathbf{y}^{n+1}\right)\right]$$

$$\le 0.5\sum_k(\mathbf{B}_k^n\cdot\mathbf{y}_k^n,\mathbf{y}_k^n)+0.5(1+c_1\Delta x)\left\|\mathbf{y}^{n+1}\right\|^2. \tag{90}$$

Since $\mathbf{y}^n(x)$ *and* $\mathbf{B}_k^n(x)$ *are periodic functions, we obtain, by virtue of (22), (89), and (90), that*

$$(1-\mu_0c_1\Delta t)\left\|\mathbf{y}^{n+1}\right\|^2\le\sum_k\left(\mathbf{B}_{k,k}^n\cdot\mathbf{y}_k^n,\mathbf{y}_k^n\right)+\sum_k\left(\left(\mathbf{B}_k^n-\mathbf{B}_{k,k}^n\right)\cdot\mathbf{y}_k^n,\mathbf{y}_k^n\right)\le$$

$$\sum_k(\mathbf{B}_k^n\cdot\mathbf{y}^n,\mathbf{y}^n)+\frac{c_2}{2m+1}\Delta x\sum_k(\mathbf{y}_k^n,\mathbf{y}_k^n)=(1+\mu_0(c_1+c_2)\Delta t)\|\mathbf{y}^n\|^2. \tag{91}$$

It follows from (91) that

$$\left\|\mathbf{y}^{n+1}\right\|^2\le\frac{1+\mu_0(c_1+c_2)\Delta t}{1-\mu_0c_1\Delta t}\|\mathbf{y}^n\|^2. \tag{92}$$

Let $\Delta t_0=const$ *such that* $1-\mu_0c_1\Delta t_0>0$. *In particular, let* $\Delta t_0=0.5\diagup(\mu_0c_1)$. *Then, for all* $\Delta t<\Delta t_0$ *the following inequality will be valid*

$$\left\|\mathbf{y}^{n+1}\right\|^2\le(1+c_3\Delta t)\|\mathbf{y}^n\|^2,\quad c_3=2\mu_0(2c_1+c_2)=const. \tag{93}$$

The inequality in (93) proves Proposition 9.

Notice, in practice, we often deal with systems for which the matrices $\mathbf{B}_k^n$ in Scheme (86) are not symmetric. Let us generalize the theorem of Friedrichs, [57, p. 120], [60, p. 374], to be applicable to variational schemes with non-symmetric matrices $\mathbf{B}_k^n$.

It is assumed that there exist matrix-valued functions $\mathbf{W}(x,t)$ and $\mathbf{A}_k(x,t)$ such that $\mathbf{W}=\mathbf{W}^T$, $\mathbf{W}\cdot\mathbf{A}_k=(\mathbf{W}\cdot\mathbf{A}_k)^T$,

$$\lambda_0(\mathbf{u},\mathbf{u})\le(\mathbf{W}\cdot\mathbf{u},\mathbf{u})\le\Lambda_0(\mathbf{u},\mathbf{u}),\quad \lambda_0,\Lambda_0=const,\ 0<\lambda_0\le\Lambda_0, \tag{94}$$

$$\left\|\mathbf{W}^{n+1}-\mathbf{W}^{n}\right\|\leq\alpha_{0}\Delta t,\quad\alpha_{0}=const,\quad\mathbf{W}^{n}=\mathbf{W}\left(x,t_{n}\right),\tag{95}$$

$$\left\|\mathbf{A}_{k}^{n}-\mathbf{B}_{k}^{n}\right\|_{\mathbf{W}^{n}}\leq\frac{\beta_{0}}{2m+1}\Delta x,\quad\beta_{0}=const,\ \mathbf{A}_{k}^{n}=\mathbf{A}_{k}\left(x,t_{n}\right),\tag{96}$$

$$\left\|\mathbf{A}_{k}^{n}-\mathbf{A}_{k,j}^{n}\right\|_{\mathbf{W}^{n}}\leq b_{0}\Delta x,\quad b_{0}=const,\ \mathbf{A}_{k,j}^{n}=\mathbf{A}_{k}\left(x+j\Delta x,t_{n}\right),\tag{97}$$

where $\|\mathbf{A}\|$ is the norm of an operator $\mathbf{A}$ on a Hilbert space equipped by the scalar product (88), while $\|\mathbf{A}\|_{\mathbf{W}}$ is the norm of an operator $\mathbf{A}$ on a Hilbert space equipped by the following scalar product

$$\left(\mathbf{u},\mathbf{v}\right)_{\mathbf{W}}\equiv\left(\mathbf{W}\cdot\mathbf{u},\mathbf{v}\right),\quad\left\|\mathbf{u}\right\|_{\mathbf{W}}\equiv\sqrt{\left(\mathbf{u},\mathbf{u}\right)_{\mathbf{W}}}.\tag{98}$$

The generalization can be addressed by the following theorem, where the notation (87), (88), and (98) will be used.

Theorem 10 *Consider an explicit scheme that can be written in the form (86), where* $\mathbf{y}^{n}\in\mathbb{R}^{M}$, $\mathbf{B}_{k}^{n}\in\mathbb{R}^{M\times M}$ *are periodic (with the period equal to 1) functions of* x. *Let* $\mathbf{W}\left(x,t\right)$ *be symmetric, i.e.* $\mathbf{W}=\mathbf{W}^{T}$, *and positive (*$\mathbf{W}>0$ *in terms of (94)) matrix-valued function, and let (96), (97) be valid for the case of symmetrizable, i.e.* $\mathbf{W}^{n}\cdot\mathbf{A}_{k}^{n}=\left(\mathbf{W}^{n}\cdot\mathbf{A}_{k}^{n}\right)^{T}$, *and periodic matrix-valued function* $\mathbf{A}_{k}^{n}\left(x\right)$. *If* $\sum_{k}\mathbf{A}_{k}\left(x,t\right)=\mathbf{I}$ *and*

$$\lambda_{1}\left(\mathbf{u},\mathbf{u}\right)\leq\left(\left(\mathbf{W}\cdot\mathbf{A}_{k}\right)\cdot\mathbf{u},\mathbf{u}\right)\leq\Lambda_{1}\left(\mathbf{u},\mathbf{u}\right),\ \lambda_{1},\Lambda_{1}=const,\ 0<\lambda_{1}\leq\Lambda_{1},\tag{99}$$

then Scheme (86) will be stable

Proof. Multiplying (86) by $\mathbf{W}^{n}$, and by $\mathbf{y}^{n+1}$, we obtain, after integrating both sides, that

$$\left\|\mathbf{y}^{n+1}\right\|_{\mathbf{W}^{n}}^{2}\leq\sum_{k}\left(\mathbf{A}_{k}^{n}\cdot\mathbf{y}_{k}^{n},\mathbf{y}^{n+1}\right)_{\mathbf{W}^{n}}+\sum_{k}\left(\left(\mathbf{B}_{k}^{n}-\mathbf{A}_{k}^{n}\right)\cdot\mathbf{y}_{k}^{n},\mathbf{y}^{n+1}\right)_{\mathbf{W}^{n}}.\tag{100}$$

Since $\mathbf{W}^{n}\cdot\mathbf{A}_{k}^{n}$ is symmetric and, in view of (99), positive, and since $\sum_{k}\mathbf{A}_{k}^{n}=\mathbf{I}$, we write:

$$\sum_{k}\left(\mathbf{A}_{k}^{n}\cdot\mathbf{y}_{k}^{n},\mathbf{y}^{n+1}\right)_{\mathbf{W}^{n}}\leq0.5\sum_{k}\left(\mathbf{A}_{k}^{n}\cdot\mathbf{y}_{k}^{n},\mathbf{y}_{k}^{n}\right)_{\mathbf{W}^{n}}+0.5\left\|\mathbf{y}^{n+1}\right\|_{\mathbf{W}^{n}}^{2}.\tag{101}$$

By virtue of (97), periodicity of $\mathbf{A}_{k}^{n}$ and $\mathbf{y}_{k}^{n}$, and since $\sum_{k}\mathbf{A}_{k}^{n}=\mathbf{I}$, we obtain:

$$\sum_{k}\left(\mathbf{A}_{k}^{n}\cdot\mathbf{y}_{k}^{n},\mathbf{y}_{k}^{n}\right)_{\mathbf{W}^{n}}=\sum_{k}\left(\mathbf{A}_{k,k}^{n}\cdot\mathbf{y}_{k}^{n},\mathbf{y}_{k}^{n}\right)_{\mathbf{W}^{n}}+\sum_{k}\left(\left(\mathbf{A}_{k}^{n}-\mathbf{A}_{k,k}^{n}\right)\cdot\mathbf{y}_{k}^{n},\mathbf{y}_{k}^{n}\right)_{\mathbf{W}^{n}}\leq$$

$$\sum_{k}\left(\mathbf{A}_{k}^{n}\cdot\mathbf{y}^{n},\mathbf{y}^{n}\right)_{\mathbf{W}^{n}}+b_{0}\Delta x\sum_{k}\left(\mathbf{y}_{k}^{n},\mathbf{y}_{k}^{n}\right)_{\mathbf{W}^{n}}\leq\left(1+b_{1}\Delta x\right)\left\|\mathbf{y}^{n}\right\|_{\mathbf{W}^{n}}^{2},\tag{102}$$

where $b_{1}=\left(2m+1\right)b_{0}=const$. In view of (96) and the assumption of periodicity, we obtain for the last term in the right-hand side of (100) that

$$\sum_{k}\left(\left(\mathbf{B}_{k}^{n}-\mathbf{A}_{k}^{n}\right)\cdot\mathbf{y}_{k}^{n},\mathbf{y}^{n+1}\right)_{\mathbf{W}^{n}}\leq\sum_{k}\left\|\left(\mathbf{B}_{k}^{n}-\mathbf{A}_{k}^{n}\right)\cdot\mathbf{y}_{k}^{n}\right\|_{\mathbf{W}^{n}}\left\|\mathbf{y}^{n+1}\right\|_{\mathbf{W}^{n}}\leq$$

$$\beta_0 \Delta x \left\|\mathbf{y}^n\right\|_{\mathbf{W}^n} \left\|\mathbf{y}^{n+1}\right\|_{\mathbf{W}^n} \leq 0.5\beta_0 \Delta x \left\|\mathbf{y}^n\right\|^2_{\mathbf{W}^n} + 0.5\beta_0 \Delta x \left\|\mathbf{y}^{n+1}\right\|^2_{\mathbf{W}^n}. \quad (103)$$

After elementary transformations we find from (100)-(103) that

$$\left\|\mathbf{y}^{n+1}\right\|^2_{\mathbf{W}^n} \leq \frac{1 + (\beta_0 + b_1)\,\Delta x}{1 - \beta_0 \Delta x} \left\|\mathbf{y}^n\right\|^2_{\mathbf{W}^n}. \quad (104)$$

Let $\Delta x_0 = const$ such that $1 - \beta_0 \Delta x_0 > 0$. In particular, let $\Delta x_0 = 0.5 \diagup \beta_0$. Then, in view of (22), for all $\Delta x < \Delta x_0$ the following inequality will be valid

$$\left\|\mathbf{y}^{n+1}\right\|^2_{\mathbf{W}^n} \leq (1 + b_2 \Delta t) \left\|\mathbf{y}^n\right\|^2_{\mathbf{W}^n},\ b_2 = 2\mu_0 \left(2\beta_0 + b_1\right) = const. \quad (105)$$

By virtue of (94) and (95), we find that

$$\left\|\mathbf{y}^{n+1}\right\|^2_{\mathbf{W}^{n+1}} = \left(\mathbf{W}^{n+1} \cdot \mathbf{y}^{n+1}, \mathbf{y}^{n+1}\right) = \left(\mathbf{W}^{n} \cdot \mathbf{y}^{n+1}, \mathbf{y}^{n+1}\right) +$$

$$\left(\left(\mathbf{W}^{n+1} - \mathbf{W}^n\right) \cdot \mathbf{y}^{n+1}, \mathbf{y}^{n+1}\right) \leq \left\|\mathbf{y}^{n+1}\right\|^2_{\mathbf{W}^n} + \alpha_0 \Delta t \left(\mathbf{y}^{n+1}, \mathbf{y}^{n+1}\right) \leq$$

$$(1 + b_3 \Delta t) \left\|\mathbf{y}^{n+1}\right\|^2_{\mathbf{W}^n}, \quad b_3 = \frac{\alpha_0}{\lambda_0} = const. \quad (106)$$

Then, in view of (105) and (106), we write

$$\left\|\mathbf{y}^{n+1}\right\|_{\mathbf{W}^{n+1}} \leq (1 + c\Delta t) \left\|\mathbf{y}^n\right\|_{\mathbf{W}^n}, \quad c = \max\left(b_2, b_3\right) = const. \quad (107)$$

The inequality in (107) proves Theorem 10. ■

3 Monotone C^1 piecewise cubics in construction of central schemes

In this section we consider some theoretical aspects for high-order interpolation and employment of monotone C^1 piecewise cubics (e.g., [16], [35]) in construction of monotone central schemes. By virtue of the operator-splitting idea (see also LOS in [59]), the following chain of equations corresponds to the problem (1)

$$\frac{1}{2}\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}\left(\mathbf{u}\right) = 0, \quad t_n < t \leq t_{n+0.5}, \quad \mathbf{u}\left(x, t_n\right) = \mathbf{u}^n\left(x\right), \quad (108)$$

$$\frac{1}{2}\frac{\partial \mathbf{u}}{\partial t} = \frac{1}{\tau}\mathbf{q}\left(\mathbf{u}\right), \quad t_{n+0.5} < t \leq t_{n+1}, \quad \mathbf{u}\left(x, t_{n+0.5}\right) = \mathbf{u}^{n+0.5}\left(x\right). \quad (109)$$

Using the central differencing, we write

$$\left.\frac{\partial \mathbf{u}}{\partial t}\right|_{t=t_{n+0.125},\ x=x_{i+0.5}} = \frac{\mathbf{u}^{n+0.25}_{i+0.5} - \mathbf{u}^n_{i+0.5}}{0.25\Delta t} + O\left(\left(\Delta t\right)^2\right), \quad (110)$$

$$\left.\frac{\partial \mathbf{f}}{\partial x}\right|_{t=t_{n+0.125},\ x=x_{i+0.5}} = \frac{\mathbf{f}^{n+0.125}_{i+1} - \mathbf{f}^{n+0.125}_{i}}{\Delta x} + O\left(\left(\Delta x\right)^2\right). \quad (111)$$

By virtue of (110)-(111) we approximate (108) on the cell $[x_i, x_{i+1}] \times [t_n, t_{n+0.25}]$ by the following difference equation

$$\mathbf{v}_{i+0.5}^{n+0.25} = \mathbf{v}_{i+0.5}^{n} - \frac{\Delta t}{2\Delta x}\left(\mathbf{g}_{i+1}^{n+0.125} - \mathbf{g}_{i}^{n+0.125}\right), \tag{112}$$

where $\mathbf{v}_i^{n+\beta}$, $\mathbf{g}_i^{n+\beta}$ are the grid functions. As usually, the mathematical treatment for the second step (i.e., on the cell $[x_{i-0.5}, x_{i+0.5}] \times [t_{n+0.25}, t_{n+0.5}]$) of a staggered scheme will, in general, not be included in the text, because it is quite similar to the one for the first step.

Considering that (112) approximate (108) with the accuracy $O((\Delta x)^2 + (\Delta t)^2)$, the next problem is to approximate $\mathbf{v}_{i+0.5}^{n}$ and $\mathbf{g}_i^{n+0.125}$ in such a way as to retain the accuracy of the approximation. For instance, the following approximations

$$\mathbf{v}_{i+0.5}^{n} = 0.5\left(\mathbf{v}_i^n + \mathbf{v}_{i+1}^n\right) + O\left((\Delta x)^2\right), \quad \mathbf{g}_i^{n+0.125} = \mathbf{f}\left(\mathbf{v}_i^n\right) + O\left(\Delta t\right), \tag{113}$$

leads to the staggered form of the famed LxF scheme that is of the first-order approximation (see, e.g., [21, p. 170]). One way to obtain a higher-order scheme is to use a higher order interpolation. At the same time it is required of the interpolant to be monotonicity preserving. Notice, the classic cubic spline does not possess such a property (see Figure 1a). Let us consider the problem of high-order interpolation of $\mathbf{v}_{i+0.5}^{n}$ in (112) with closer inspection

Let $\mathbf{p} = \mathbf{p}(x) \equiv \left\{p^1(x), \ldots, p^k(x), \ldots, p^m(x)\right\}^T$ be a component-wise monotone C^1 piecewise cubic interpolant (e.g., [16], [35]), and let

$$\mathbf{p}_i = \mathbf{p}(x_i), \quad \mathbf{p}_i' = \mathbf{p}'(x_i), \quad \Delta\mathbf{p}_i = \mathbf{p}_{i+1} - \mathbf{p}_i,$$

$$\mathbf{p}_i' = \mathbb{A}_i \cdot \frac{\Delta\mathbf{p}_i}{\Delta x}, \quad \mathbf{p}_{i+1}' = \mathbb{B}_i \cdot \frac{\Delta\mathbf{p}_i}{\Delta x}, \tag{114}$$

where $\mathbf{p}_i'$ denotes the derivative of the interpolant at $x = x_i$. The diagonal matrices $\mathbb{A}_i$ and $\mathbb{B}_i$ in (114) are defined as follows

$$\mathbb{A}_i = diag\left\{\alpha_i^1, \alpha_i^2, \ldots, \alpha_i^m\right\}, \ \mathbb{B}_i = diag\left\{\beta_i^1, \beta_i^2, \ldots, \beta_i^m\right\}. \tag{115}$$

The cubic interpolant, $\mathbf{p} = \mathbf{p}(x)$, is component-wise monotone on $[x_i, x_{i+1}]$ *iff* one of the following conditions (e.g., [16], [35]) is satisfied:

$$\left(\alpha_i^k - 1\right)^2 + \left(\alpha_i^k - 1\right)\left(\beta_i^k - 1\right) + \left(\beta_i^k - 1\right)^2 - 3\left(\alpha_i^k + \beta_i^k - 2\right) \le 0, \tag{116}$$

$$\alpha_i^k + \beta_i^k \le 3, \quad \alpha_i^k \ge 0, \ \beta_i^k \ge 0, \quad \forall i, k. \tag{117}$$

As reported in [35], the necessary and sufficient conditions for monotonicity of a C^1 piecewise cubic interpolant originally given by Ferguson and Miller (1969), and independently, by Fritsch and Carlson [16]. The region of monotonicity is shown in Figure 1b. The results of implementing a monotone C^1 piecewise cubic interpolation when compared with the classic cubic spline interpolation,

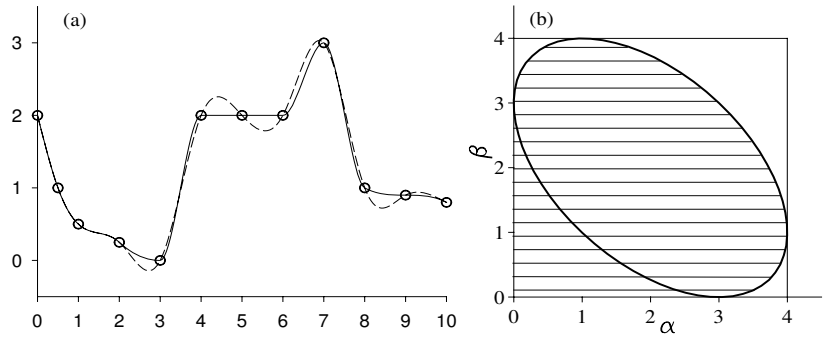

Figure 1: Monotone piecewise cubic interpolation. (a) Interpolation of a 1-D tabulated function. Circles: prescribed tabulated values; Dashed line: classic cubic spline; Solid line: monotone piecewise cubic. (b) Necessary and sufficient conditions for monotonicity. Horizontal hatching: region of monotonicity; Unshaded: cubic is non-monotone.

are depicted in Figure 1a. We note (Figure 1a) that the constructed function produces monotone interpolation and this function coincides with the classic cubic spline at some sections where the classic cubic spline is monotone.

Using the cubic segment of the C^1 piecewise cubic interpolant, $\mathbf{p} = \mathbf{p}(x)$, (see, e.g., [16], [35]) for $x \in [x_i, x_{i+1}]$, we obtain the following interpolation formula

$$\mathbf{p}_{i+0.5} = 0.5\left(\mathbf{p}_i + \mathbf{p}_{i+1}\right) - \frac{\Delta x}{8}\left(\mathbf{p}'_{i+1} - \mathbf{p}'_i\right) + O\left((\Delta x)^r\right). \tag{118}$$

If $\mathbf{p}(x)$ has a continuous fourth derivative, then $r = 4$ in (118), see e.g. [33, p. 111]. However, the exact value of $\mathbf{p}'_i$ in (118) is, in general, unknown, and hence to construct numerical schemes, employing formulae similar to (118), the value of derivatives $\mathbf{p}'_i$ must be estimated.

Using (118) and the second formula in (113) we obtain from (112) the following scheme

$$\mathbf{v}_{i+0.5}^{n+0.25} = 0.5\left(\mathbf{v}_i^n + \mathbf{v}_{i+1}^n\right) - \frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^n - \mathbf{d}_i^n\right) - \frac{\Delta t}{2}\frac{\mathbf{f}\left(\mathbf{v}_{i+1}^n\right) - \mathbf{f}\left(\mathbf{v}_i^n\right)}{\Delta x}, \tag{119}$$

where $\mathbf{d}_i^n$ denotes the derivative of the interpolant at $x = x_i$. In view of (118) and the second formula in (113), the local truncation error [40, p. 142], ψ, on a sufficiently smooth solution $\mathbf{u}(x,t)$ to (108) is found to be

$$\psi = O\left(\Delta t\right) + O\left(\frac{\left(\Delta x\right)^r}{\Delta t}\right) + O\left(\left(\Delta t\right)^2 + \left(\Delta x\right)^2\right). \tag{120}$$

In view of (120) we conclude that the scheme (119) generates a conditional approximation, because it approximates (108) only if $\left(\Delta x\right)^r / \Delta t \to 0$ as $\Delta x \to 0$ and $\Delta t \to 0$. Let $\mathbf{d}_i^n$ be approximated with the accuracy $O\left((\Delta x)^s\right)$, then

the value of r in (120) can be calculated (see Section 6, Proposition 11) by the following formula

$$r = \min(4, s+1). \tag{121}$$

Interestingly, since (119) provides the conditional approximation, the order of accuracy depends on the pathway taken by Δx and Δt as $\Delta x \to 0$ and $\Delta t \to 0$. Actually, there exists a pathway such that Δt is proportional to $(\Delta x)^{\mu}$ and the CFL condition is fulfilled provided $\mu \geq 1$ and $\Delta x \leq \Delta x_0$, where Δx_0 is a positive value. If we take $\mu = 1$ and $s \geq 1$, then we obtain from (120) that the scheme (119) is of the first-order. If $\mu = 2$ and $s \geq 3$, then (119) is of the second-order. However, if $\mu = 2$ and $s = 2$, then, in view of (120) and (121), the scheme (119) is of the first-order. Moreover, under $\mu = 2$ and $s = 2$, the scheme will be of the first-order even if $\mathbf{g}_i^{n+0.125}$ in (112) will be approximated with the accuracy $O((\Delta t)^2)$. It seems likely that Example 6 in [37] can be seen as an illustration of the last assertion. The Nessyahu-Tadmor (NT) scheme with the second-order approximation of $\mathbf{d}_i^n$ is used [37] to solve a Burgers-type equation. Since $\Delta t = O((\Delta x)^2)$ [37], the NT scheme is of the first-order, and hence it can be the main reason for the scheme to exhibit the smeared discontinuity computed in [37, Fig. 6.22].

The approximation of derivatives $\mathbf{p}_i'$ can be done by the following three steps [16]: (i) an initialization of the derivatives $\mathbf{p}_i'$; (ii) the choice of subregion of monotonicity; (iii) modification of the initialized derivatives $\mathbf{p}_i'$ to produce a monotone interpolant.

The matter of initialization of the derivatives is the most subtle issue of this algorithm. Actually, the approximation of $\mathbf{p}_i'$ must, in general, be done with accuracy $O((\Delta x)^3)$ to obtain the second-order scheme when Δt is proportional to $(\Delta x)^2$, inasmuch as central schemes generate a conditional approximation. Thus, using the two-point or the three-point (centered) difference formula (e.g. [35], [51]) we obtain, in general, the first-order scheme. The so called limiter functions [35] lead, in general, to a low-order scheme as these limiters are often $O(\Delta x)$ or $O((\Delta x)^2)$ accurate. Performing the initialization of the derivatives $\mathbf{p}_i'$ in the interpolation formula (118) by the classic cubic spline interpolation [56], we obtain the approximation, which is $O((\Delta x)^3)$ accurate (e.g., [33], [35]), and hence, in general, the second-order scheme. The same accuracy, $O((\Delta x)^3)$, can be achieved by using the four-point approximation [35]. However, the efficiency of the algorithm based on the classic cubic spline interpolation is comparable with the one based on the four-point approximation, as the number of multiplications and divisions (as well as additions and subtractions) per one node is approximately the same for both algorithms. We will use the classic cubic spline interpolation for the initialization of the derivatives $\mathbf{P}_i'$ in the interpolation formula (118), as it is based on the tridiagonal algorithm, which is 'the rare case of an algorithm that, in practice, is more robust than theory says it should be' [56].

Obviously, for each interval $[x_i, x_{i+1}]$ in which the initialized derivatives $\mathbf{p}_i'$, $\mathbf{p}_{i+1}'$ such that at least one point $(\alpha_i^k,\ \beta_i^k)$ does not belong to the region of monotonicity (116)-(117), the derivatives $\mathbf{p}_i'$, $\mathbf{p}_{i+1}'$ must be modified to $\widetilde{\mathbf{p}}_i'$, $\widetilde{\mathbf{p}}_{i+1}'$

such that the point $(\widetilde{\alpha}_i^k, \widetilde{\beta}_i^k)$ will be in the region of monotonicity. The modification of the initialized derivatives, would be much simplified if we take a square as a subregion of monotonicity. In connection with this, we will make use the subregions of monotonicity represented in the following form:

$$0 \le \alpha_i^k \le 4\aleph, \quad 0 \le \beta_i^k \le 4\aleph, \quad \forall i, k, \tag{122}$$

where $\aleph$ is a monotonicity parameter. Obviously, the condition (122) is sufficient for the monotonicity (see Figure 1b) provided that $0 \le \aleph \le 0.75$.

Let us now find necessary and sufficient conditions for (118) to be G-monotone. By virtue of (114), the interpolation formula (118) can be rewritten to read

$$\mathbf{p}_{i+0.5} = \left(0.5\mathbf{I} + \frac{\mathbb{B}_i - \mathbb{A}_i}{8}\right) \cdot \mathbf{p}_i + \left(0.5\mathbf{I} - \frac{\mathbb{B}_i - \mathbb{A}_i}{8}\right) \cdot \mathbf{p}_{i+1}. \tag{123}$$

The coefficients of (123) will be non-negative *iff* $|\beta_i - \alpha_i| \le 4$. Hence (118) will be G-monotone *iff* (122) will be valid provided $0 \le \aleph \le 1$. Notice, there is no any contradiction between the sufficient conditions, (122) provided $0 \le \aleph \le 0.75$, for the interpolant, $\mathbf{p} = \mathbf{p}(x)$, to be monotone through the interval $[x_i, x_{i+1}]$, and the necessary and sufficient conditions, (122) provided $0 \le \aleph \le 1$, for the scheme (123) to be G-monotone. In the latter case the interpolant, $\mathbf{p} = \mathbf{p}(x)$, may, in general, be non-monotone, however at the point $i + 0.5$ the value of an arbitrary component of $\mathbf{p}_{i+0.5}$ will be between the corresponding components of $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$.

To fulfill the conditions of monotonicity (122), the modification of derivatives $\mathbf{p}_i' = \left\{p_i'^1, p_i'^2, \ldots, p_i'^m\right\}$ can be done by the following algorithm suggested, in fact, by Fritsch and Carlson [16] (see also [35]):

$$S_i^k := 4\aleph \min\mathrm{mod}(\Delta_{i-1}^k, \Delta_i^k), \quad \widetilde{p}_i'^k := \min\mathrm{mod}(p_i'^k, S_i^k), \quad \aleph = const, \tag{124}$$

where $\Delta_i^k = \left(p_{i+1}^k - p_i^k\right) \diagup \Delta x$, the function $\min\mathrm{mod}(x, y)$ is defined (e.g., [35], [37], [45], [51], [61]) as follows

$$\min\mathrm{mod}(x, y) \equiv \frac{1}{2}\left[sgn(x) + sgn(y)\right] \min\left(|x|, |y|\right). \tag{125}$$

4 Construction of first- and second-order central schemes

Central difference schemes with first- and second-order accuracy are introduced in this section. The construction of the central schemes is based on: (i) Variational GOS-monotonicity notion, (ii) Monotone piecewise cubic interpolation (e.g., [16], [35]), (iii) Operator-splitting techniques (see also LOS in [59]).

4.1 First-order central schemes

Let us note that instead of point values, $\mathbf{v}_{i+0.5}^n$, employed in the construction of the scheme (112), it can be used the cell averages (e.g., [6], [37], [40]) calculated

on the basis of the monotone C^1 piecewise cubics. In such a case we obtain, instead of (118), the following interpolation formula

$$\mathbf{p}_{i+0.5} = 0.5\left(\mathbf{p}_i + \mathbf{p}_{i+1}\right) - \varkappa\frac{\Delta x}{8}\left(\mathbf{p}'_{i+1} - \mathbf{p}'_i\right), \tag{126}$$

where $\varkappa = 2/3$. The region of monotonicity in this case will also be

$$0 \le \mathbb{A}_i \le 4\aleph\mathbf{I},\ 0 \le \mathbb{B}_i \le 4\aleph\mathbf{I}, \quad 0 \le \aleph \le 1, \quad \forall i. \tag{127}$$

Notice, the interpolation formula (126) coincides with (118) under $\varkappa = 1$. Thus, in view of the interpolation formula (126), the staggered scheme (112) is written to read

$$\mathbf{v}_{i+0.5}^{n+0.25} = 0.5\left(\mathbf{v}_{i+1}^{n} + \mathbf{v}_i^{n}\right) - \varkappa\frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^{n} - \mathbf{d}_i^{n}\right) - \frac{\Delta t}{2}\frac{\mathbf{f}\left(\mathbf{v}_{i+1}^{n}\right) - \mathbf{f}\left(\mathbf{v}_i^{n}\right)}{\Delta x}, \tag{128}$$

where $\mathbf{d}_i^n$ denotes the derivative of the interpolant at $x = x_i$, the range of values for the parameter $\varkappa$ is the segment $0 \le \varkappa \le 1$. As usually, the mathematical treatments for the second step of the staggered scheme will, in general, not be included in the text, because the second step is quite similar to (128). If $\varkappa = 1$ (or $\varkappa = 0$), then Scheme (128) coincides with the scheme (119) (or with the LxF scheme, respectively). As it was shown above, the scheme (128) is of the first order provided $\Delta t = O(\Delta x)$. In such a case, since the source terms can be, in general, stiff (i.e., $\tau \ll 1$), it is natural to use the following first-order implicit scheme for (109).

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^{n+0.5} + \frac{\Delta t}{\tau}\mathbf{q}\left(\mathbf{v}_i^{n+1}\right). \tag{129}$$

The first order central scheme (128)-(129) based on operator-splitting techniques will be abbreviated to as COS1.

Let us investigate the stability of Scheme (128). It is assumed that the vector-valued function $\mathbf{u} = \mathbf{u}(x,t)$ is Gâteaux- (or Fréchet-) differentiable on the convex set $\Omega_{xt} \subset \mathbb{R} \times [0, +\infty)$, and its derivative is bounded on Ω_{xt}. It is also assumed that the operator $\mathbf{A}$ $(= \partial\mathbf{f}(\mathbf{u})/\partial\mathbf{u})$ is Fréchet-differentiable on the convex set $\Omega_{\mathbf{u}} \subset \mathbb{R}^M$, and its derivative is bounded on $\Omega_{\mathbf{u}}$. Hence $\mathbf{A} = \mathbf{A}(x,t)$ will be Gâteaux- (or, respectively, Fréchet-) differentiable [49, p. 62] and its derivative will be bounded on Ω_{xt}, and hence $\mathbf{A}(x,t)$ will be Lipschitz-continuous on Ω_{xt} [49, p. 70]. Since the Jacobian matrix $\mathbf{A} = \mathbf{A}(x,t)$ in (2) possesses M linearly independent eigenvectors, $\mathbf{A}_i^n$ $(= \mathbf{A}(x_i, t_n))$ is similar to a diagonal matrix [43], i.e. there exists a non-singular matrix $\mathbf{S}_i^n = \mathbf{S}(x_i, t_n)$ such that

$$\left(\mathbf{S}_i^n\right)^{-1} \cdot \mathbf{A}_i^n \cdot \mathbf{S}_i^n = diag\left\{\lambda_i^{n,1}, \lambda_i^{n,2}, \ldots, \lambda_i^{n,M}\right\}, \quad \forall i, n. \tag{130}$$

The right and left eigenvectors of $\mathbf{A}_i^n = \mathbf{A}(x_i, t_n)$ can be defined in such a way that $\mathbf{S}_i^n = \mathbf{S}(x_i, t_n)$ will be the matrix having the right eigenvectors as its columns, and the rows of $\left(\mathbf{S}_i^n\right)^{-1} = \mathbf{S}^{-1}(x_i, t_n)$ will be the left eigenvectors [39, p. 62]. For Theorem 7 to be used, it must be proven that $\left[\left(\mathbf{S}_i^n\right)^{-1} -\right.$

$\left(\mathbf{S}^n\right)^{-1}(x)\Big] \cdot \mathbf{S}^n(x)$ and $\left[\left(\mathbf{S}_i\right)^{-1}(t) - \left(\mathbf{S}_i^n\right)^{-1}\right] \cdot \mathbf{S}_i^n$ will be Lipschitz-continuous in space and, respectively, time $\forall i, n$. Since each of the above functions depends on one parameter (x or t) only, it can be done with ease for strictly hyperbolic systems, i.e. when the eigenvalues of the operator $\mathbf{A}$ ($= \partial \mathbf{f}(\mathbf{u}) \diagup \partial \mathbf{u}$) in (2) will be all distinct [40, p. 2]. In this case the proof follows from the long-known results of perturbation theory for simple eigenvalues [62, p. 67] (see also Theorem 2.1 in [4]). For a detailed discussion on the derivatives (sensitivities) of eigenvectors of matrix-valued functions depending on several parameters when the eigenvalues are multiple see [4], [63].

By virtue of (114), the second term in right-hand side of (128) can be written in the form

$$\varkappa \frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^n - \mathbf{d}_i^n\right) = \frac{\varkappa}{8}\left(\mathbb{B}_i^n - \mathbb{A}_i^n\right) \cdot \left(\mathbf{v}_{i+1}^n - \mathbf{v}_i^n\right). \tag{131}$$

Then, the variational scheme corresponding to (128) is the following

$$\delta \mathbf{v}_{i+0.5}^{n+0.25} = 0.5\left(\delta \mathbf{v}_i^n + \delta \mathbf{v}_{i+1}^n\right) + \frac{\varkappa}{8}\left[\left(\mathbf{v}_i^n - \mathbf{v}_{i+1}^n\right)^T \cdot \delta \mathbb{D}_i^n\right]^T +$$

$$\frac{\varkappa}{8}\mathbb{D}_i^n \cdot \left(\delta \mathbf{v}_i^n - \delta \mathbf{v}_{i+1}^n\right) + \frac{\Delta t}{2\Delta x}\left(\mathbf{A}_i^n \cdot \delta \mathbf{v}_i^n - \mathbf{A}_{i+1}^n \cdot \delta \mathbf{v}_{i+1}^n\right), \tag{132}$$

where $\mathbb{D}_i^n = diag\left\{D_{i,1}^n, D_{i,2}^n, \ldots, D_{i,M}^n\right\} \equiv \mathbb{B}_i^n - \mathbb{A}_i^n$. By virtue of (127), we find that $-4\aleph\mathbf{I} \leq \mathbb{D}_i^n \leq 4\aleph\mathbf{I}$, and hence $-8\aleph\mathbf{I} \leq \delta\mathbb{D}_i^n \leq 8\aleph\mathbf{I}$. Thus, we may write that

$$\left\|\delta \mathbb{D}_i^n\right\|_2 \leq 8\aleph. \tag{133}$$

Considering that $\mathbf{v}_i^n$ in (128) is Lipschitz-continuous on Ω_{xt}, we write

$$\left\|\mathbf{v}_i^n - \mathbf{v}_{i+1}^n\right\|_2 \leq C_v \Delta x, \quad C_v = const. \tag{134}$$

It is assumed, see (22), that there exists $\alpha_0 = const$ such that $\Delta x \leq \alpha_0 \Delta t$ for a sufficiently small Δt. Then, by virtue of (133) and (134), and since $0 \leq \varkappa, \aleph \leq 1$, we find the following estimation for the second term in right-hand side of (132):

$$\left\|\frac{\varkappa}{8}\left[\left(\mathbf{v}_i^n - \mathbf{v}_{i+1}^n\right)^T \cdot \delta \mathbb{D}_i^n\right]^T\right\|_2 \leq \frac{\varkappa}{8}\left\|\mathbf{v}_i^n - \mathbf{v}_{i+1}^n\right\|_2 \left\|\delta \mathbb{D}_i^n\right\|_2 \leq \alpha_0 C_v \Delta t. \tag{135}$$

In view of (135), the scheme (132) will be stable if the following scheme will be stable (e.g., [59, pp. 390-392])

$$\delta \mathbf{v}_{i+0.5}^{n+0.25} = 0.5\left(\mathbf{I} + \mathbf{E}_i^n\right) \cdot \delta \mathbf{v}_i^n + 0.5\left(\mathbf{I} - \mathbf{E}_{i+1}^n\right) \cdot \delta \mathbf{v}_{i+1}^n \equiv$$

$$0.5\left(\mathbf{I} + \frac{\varkappa}{4}\mathbb{D}_i^n + \frac{\Delta t}{\Delta x}\mathbf{A}_i^n\right) \cdot \delta \mathbf{v}_i^n + 0.5\left(\mathbf{I} - \frac{\varkappa}{4}\mathbb{D}_i^n - \frac{\Delta t}{\Delta x}\mathbf{A}_{i+1}^n\right) \cdot \delta \mathbf{v}_{i+1}^n. \tag{136}$$

We write, in view of (2) and (127), that

$$s\left(\mathbf{E}_i^n\right) \subset \left[-\varkappa\aleph - \frac{\Delta t}{\Delta x}\lambda_{\max},\ \varkappa\aleph + \frac{\Delta t}{\Delta x}\lambda_{\max}\right], \quad \forall i, n. \tag{137}$$

Hence, by virtue of Theorem 7 we find that the scheme (136) will be stable if

$$\max_{\lambda\in s\left(\mathbf{E}_i^n\right)} 0.5\left(|1+\lambda|+|1-\lambda|\right)\leqslant 1,\quad \forall i,n. \tag{138}$$

We obtain from (138) the following condition for the stability of the variational scheme (136):

$$\varkappa\aleph + C_r \leqslant 1,\quad C_r=\frac{\Delta t\lambda_{\max}}{\Delta x}, \tag{139}$$

where C_r denotes the Courant number. Thus, in view of Theorem 3 the scheme (128) will be stable if (139) will be valid.

Notice, (139) was obtained on the basis that $\mathbf{E}_i^n\equiv\frac{\varkappa}{4}\mathbb{D}_i^n+\frac{\Delta t}{\Delta x}\mathbf{A}_i^n$ is diagonizable. Such an assumption should be verified for a concrete problem. It might be well to point out that this assumption has a rigorous basis if $\mathbf{A}_i^n$ is diagonizable and $\mathbb{D}_i^n$ is a scalar matrix, i.e. $\mathbb{D}_i^n=4\theta^n\mathbf{I}$, $\theta^n=const$ $(-\aleph\leq\theta^n\leq\aleph)$. In such a case $\mathbf{E}_i^n$ will be diagonizable. Thus, for instance, LxF scheme will be stable if $C_r\leqslant 1$. In the case, when $\mathbb{D}_i^n$ is not a scalar matrix, whereas the Jacobian matrix $\mathbf{A}$ in (2) is symmetric, the condition (139) for stability of (128) can be found with ease by virtue of Proposition 9. In a more general case, when $\mathbf{E}_i^n$ is not diagonizable as well as $\mathbf{A}$ is not symmetric, the stability of (128) can be investigated by virtue of Theorem 10.

Let us find a necessary condition for the variational S-monotonicity (see Definition 1) of the COS1 scheme, (128)-(129). Considering (132) on a constant $(\mathbf{v}_i^n=\mathbf{C}=const,\ \forall i,n)$, we obtain

$$\delta\mathbf{v}_{i+0.5}^{n+0.25}=0.5\left(\delta\mathbf{v}_i^n+\delta\mathbf{v}_{i+1}^n\right)+\frac{\varkappa}{8}\mathbb{D}\cdot\left(\delta\mathbf{v}_i^n-\delta\mathbf{v}_{i+1}^n\right)+$$

$$+\frac{\Delta t}{2\Delta x}\mathbf{A}_c\cdot\left(\delta\mathbf{v}_i^n-\delta\mathbf{v}_{i+1}^n\right),\quad \mathbf{A}_c=\mathbf{A}\left(\mathbf{C}\right),\quad \mathbf{A}\left(\mathbf{u}\right)=\frac{\partial\mathbf{f}\left(\mathbf{u}\right)}{\partial\mathbf{u}}. \tag{140}$$

Here, in (140), $\mathbf{A}_c$ is a diagonizable matrix such that its spectrum $s\left(\mathbf{A}_c\right)\subset[-\lambda_{\max},\ \lambda_{\max}]$, see (2). In view of (131), $\mathbb{D}$ in (140) can be taken at will, as $\mathbf{v}_i^n=\mathbf{C}=const$, $\mathbf{v}_{i+1}^n-\mathbf{v}_i^n=0$ and $\mathbf{d}_i^n=\mathbf{d}_{i+1}^n=0$. Aiming to obtain the necessary condition for (132) to be S-monotone, it is assumed that $\mathbb{D}=\zeta\mathbf{I}$ is a scalar matrix with $\zeta\in[-4\aleph,4\aleph]$. In view of Theorem 8, the scheme (140) will be S-monotone *iff*

$$\max_{|\zeta|\leq 4\aleph,\ \lambda\in s(\mathbf{A}_c)}\left|\frac{1}{2}+\frac{\varkappa\zeta}{8}+\frac{\Delta t}{2\Delta x}\lambda\right|+\left|\frac{1}{2}-\frac{\varkappa\zeta}{8}-\frac{\Delta t}{2\Delta x}\lambda\right|\leqslant 1. \tag{141}$$

By virtue of (141), we find that (139) is the necessary condition for the variational S-monotonicity of (128).

The variational scheme corresponding to (129) reads

$$\delta\mathbf{v}_i^{n+1}=\delta\mathbf{v}_i^{n+0.5}+\frac{\Delta t}{\tau}\mathbf{G}\left(\mathbf{v}_i^{n+1}\right)\cdot\delta\mathbf{v}_i^{n+1}, \tag{142}$$

where $\mathbf{G}\left(\mathbf{V}\right)=\partial\mathbf{q}\left(\mathbf{V}\right)\diagup\partial\mathbf{V}$. Let us rewrite (142) to read

$$\delta\mathbf{v}_i^{n+1}=\left\{\mathbf{I}-\frac{\Delta t}{\tau}\mathbf{G}\left(\mathbf{v}_i^{n+1}\right)\right\}^{-1}\cdot\delta\mathbf{v}_i^{n+0.5}. \tag{143}$$

Let $\mathbf{G}$ will be a normal matrix. In such a case, in view of [10, Theorem 3.3] the variational scheme (143) will be S-monotone if

$$\max_k \frac{1}{\left|1-\frac{\Delta t}{\tau}\xi_k\left(\mathbf{v}_i^{n+1}\right)\right|} \leq 1, \quad \forall i,n, \tag{144}$$

where ξ_k denotes the $k-th$ eigenvalue of the matrix $\mathbf{G}$. In view of (3), the inequality (144) is valid, and hence the variational scheme (142) will be unconditionally S-monotone.

Thus, the COS1 scheme, (128)-(129), will be variationally S-monotone only if (139) will be valid.

4.2 Second-order central scheme

In this section, the second-order scheme for (108) is developed by approximating $\mathbf{v}_{i+0.5}^n$ and $\mathbf{g}_i^{n+0.125}$ in (112) with the accuracy $O((\Delta x)^2+(\Delta t)^2)$. The sufficient conditions for stability as well as the necessary condition for S-monotonicity of this scheme are found.

Using Taylor series expansion, we write

$$\mathbf{g}_i^{n+0.125} = \mathbf{f}\left(\mathbf{v}_i^n\right) + \left.\frac{\partial \mathbf{f}\left(\mathbf{v}_i^n\right)}{\partial t}\right|_{t=t_n} \frac{\Delta t}{8} + O\left(\Delta t^2\right). \tag{145}$$

By virtue of the PDE system, (108), we find

$$\frac{\partial \mathbf{f}}{\partial t} = \frac{\partial \mathbf{f}}{\partial \mathbf{u}}\cdot\frac{\partial \mathbf{u}}{\partial t} = -2\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\cdot\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\cdot\frac{\partial \mathbf{u}}{\partial x} = -2\left(\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right)^2\cdot\frac{\partial \mathbf{u}}{\partial x}. \tag{146}$$

Using the interpolation formula (126) and the formulae (145)-(146), we obtain from (112) the following second order central scheme

$$\mathbf{v}_{i+0.5}^{n+0.25} = 0.5\left(\mathbf{v}_{i+1}^n+\mathbf{v}_i^n\right) - \varkappa\frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^n-\mathbf{d}_i^n\right) +$$

$$\frac{(\Delta t)^2}{8\Delta x}\left[\left(\mathbf{A}_{i+1}^n\right)^2\cdot\mathbf{d}_{i+1}^n - \left(\mathbf{A}_i^n\right)^2\cdot\mathbf{d}_i^n\right] - \frac{\Delta t}{2}\frac{\mathbf{f}\left(\mathbf{v}_{i+1}^n\right)-\mathbf{f}\left(\mathbf{v}_i^n\right)}{\Delta x}. \tag{147}$$

Since $\mathbf{d}_i^n$ is the derivative of the interpolant at $x=x_i$, the third term in the right-hand side of (147) can be seen as the non-negative numerical viscosity introduced into the first order scheme (128). Owing to this term, the scheme (147) is $O((\Delta x)^2+(\Delta t)^2)$ accurate. Since the source terms in (1) can be, in general, stiff (i.e., $\tau \ll 1$), it is natural to use the following second-order implicit Runge-Kutta scheme for (109), since this scheme possesses a discrete analogy to the continuous asymptotic limit,

$$\mathbf{v}_i^{n+0.75} = \mathbf{v}_i^{n+1} - \frac{\gamma}{2}\mathbf{q}\left(\mathbf{v}_i^{n+1}\right),$$

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^{n+0.5} + \gamma\mathbf{q}\left(\mathbf{v}_i^{n+0.75}\right), \quad \gamma \equiv \frac{\Delta t}{\tau}. \tag{148}$$

Scheme (147)-(148), based on operator-splitting techniques, will be abbreviated to as COS2.

Let us find the sufficient conditions for stability of Scheme (147). It is assumed that the following inequalities are valid in a suitable norm

$$\left\| \frac{\partial \mathbf{A}_i^n}{\partial \mathbf{v}_i^n} \cdot \delta \mathbf{v}_i^n \right\| \leq \beta_A \left\| \delta \mathbf{v}_i^n \right\|, \quad \alpha_A, \beta_A = const. \tag{149}$$

In view of (114), the variational scheme corresponding to (147) is the following

$$\delta \mathbf{v}_{i+0.5}^{n+0.25} = 0.5 \left(\delta \mathbf{v}_i^n + \delta \mathbf{v}_{i+1}^n \right) + \frac{\varkappa}{8} \left[\left(\mathbf{v}_i^n - \mathbf{v}_{i+1}^n \right)^T \cdot \delta \mathbb{D}_i^n \right]^T + \frac{\varkappa}{8} \mathbb{D}_i^n \cdot \left(\delta \mathbf{v}_i^n - \delta \mathbf{v}_{i+1}^n \right)$$

$$+ \frac{(\Delta t)^2}{8 \Delta x^2} \left\{ \left[\delta \left(\left(\mathbf{A}_{i+1}^n \right)^2 \cdot \mathbb{B}_i \right) \right] \cdot \left(\mathbf{v}_{i+1}^n - \mathbf{v}_i^n \right) - \left[\delta \left(\left(\mathbf{A}_i^n \right)^2 \cdot \mathbb{A}_i \right) \right] \cdot \left(\mathbf{v}_{i+1}^n - \mathbf{v}_i^n \right) \right\} +$$

$$\frac{(\Delta t)^2}{8 \Delta x^2} \left[\left(\mathbf{A}_{i+1}^n \right)^2 \cdot \mathbb{B}_i - \left(\mathbf{A}_i^n \right)^2 \cdot \mathbb{A}_i \right] \cdot \left(\delta \mathbf{v}_{i+1}^n - \delta \mathbf{v}_i^n \right) +$$

$$\frac{\Delta t}{2 \Delta x} \left(\mathbf{A}_i^n \cdot \delta \mathbf{v}_i^n - \mathbf{A}_{i+1}^n \cdot \delta \mathbf{v}_{i+1}^n \right), \tag{150}$$

In view of (2), (133), (134), and (149), Scheme (150) will be stable if the following scheme will be stable.

$$\delta \mathbf{v}_{i+0.5}^{n+0.25} = 0.5 \left(\delta \mathbf{v}_i^n + \delta \mathbf{v}_{i+1}^n \right) + \frac{\varkappa}{8} \mathbb{D}_i^n \cdot \left(\delta \mathbf{v}_i^n - \delta \mathbf{v}_{i+1}^n \right) +$$

$$\frac{(\Delta t)^2}{8 \Delta x^2} \left[\left(\mathbf{A}_{i+1}^n \right)^2 \cdot \mathbb{B}_i - \left(\mathbf{A}_i^n \right)^2 \cdot \mathbb{A}_i \right] \cdot \left(\delta \mathbf{v}_{i+1}^n - \delta \mathbf{v}_i^n \right) +$$

$$\frac{\Delta t}{2 \Delta x} \left(\mathbf{A}_i^n \cdot \delta \mathbf{v}_i^n - \mathbf{A}_{i+1}^n \cdot \delta \mathbf{v}_{i+1}^n \right). \tag{151}$$

We rewrite (151) to read

$$\delta \mathbf{v}_{i+0.5}^{n+0.25} = 0.5 \left(\mathbf{I} + \mathbf{E}_i^n \right) \cdot \delta \mathbf{v}_i^n + 0.5 \left(\mathbf{I} - \mathbf{E}_{i+1}^n \right) \cdot \delta \mathbf{v}_{i+1}^n, \tag{152}$$

where

$$\mathbf{E}_i^n = \frac{\varkappa}{4} \mathbb{D}_i^n - \frac{(\Delta t)^2}{4 (\Delta x)^2} \left[\left(\mathbf{A}_{i+1}^n \right)^2 \cdot \mathbb{B}_i - \left(\mathbf{A}_i^n \right)^2 \cdot \mathbb{A}_i \right] + \frac{\Delta t}{\Delta x} \mathbf{A}_i^n, \tag{153}$$

$$\mathbf{E}_{i+1}^n = \frac{\varkappa}{4} \mathbb{D}_i^n - \frac{(\Delta t)^2}{4 (\Delta x)^2} \left[\left(\mathbf{A}_{i+1}^n \right)^2 \cdot \mathbb{B}_i - \left(\mathbf{A}_i^n \right)^2 \cdot \mathbb{A}_i \right] + \frac{\Delta t}{\Delta x} \mathbf{A}_{i+1}^n. \tag{154}$$

We write, in view of (2) and (127), that $s \left(\mathbf{E}_i^n \right) \subset [-\lambda_E, \lambda_E]$, where

$$\lambda_E = \varkappa \aleph - \left(\frac{\Delta t \lambda_{\max}}{\Delta x} \right)^2 \aleph + \frac{\Delta t}{\Delta x} \lambda_{\max}, \quad \forall i, n. \tag{155}$$

Hence, by virtue of Theorem 7 we find that the scheme (152) will be stable if

$$\left(\varkappa - C_r^2 \right) \aleph + C_r \leq 1, \quad C_r = \frac{\Delta t \lambda_{\max}}{\Delta x}. \tag{156}$$

Thus, in view of Theorem 3, the scheme (147) will be stable if (156) will be valid.

By analogy with the COS1 scheme, we find the necessary conditions for the variational S-monotonicity (see Definition 1) of the COS2 scheme, (147)-(148), assuming that $\mathbf{v}_i^n = \mathbf{C} = const$. The variational scheme corresponding to (147) under (114), (131) is the following

$$\delta\mathbf{v}_{i+0.5}^{n+0.25} = 0.5\left(\delta\mathbf{v}_i^n + \delta\mathbf{v}_{i+1}^n\right) + \frac{\varkappa}{8}\mathbb{D}\cdot\left(\delta\mathbf{v}_i^n - \delta\mathbf{v}_{i+1}^n\right) -$$

$$\frac{(\Delta t)^2}{8\left(\Delta x\right)^2}\left(\mathbf{A}_c\right)^2\cdot\mathbb{D}\cdot\left(\delta\mathbf{v}_i^n - \delta\mathbf{v}_{i+1}^n\right) + \frac{\Delta t}{2\Delta x}\mathbf{A}_c\cdot\left(\delta\mathbf{v}_i^n - \delta\mathbf{v}_{i+1}^n\right), \tag{157}$$

where $\mathbf{A}_c = \mathbf{A}\left(\mathbf{C}\right)$, $\mathbf{A}\left(\mathbf{u}\right) = \partial\mathbf{f}\left(\mathbf{u}\right)/\partial\mathbf{u}$, $\mathbb{D} = \zeta\mathbf{I}$. In view of Theorem 8, the scheme (157) will be S-monotone *iff*

$$\max_{|\zeta|\leq 4\aleph,\ \lambda\in s(\mathbf{A}_c)}\left(\left|\frac{1}{2} + \frac{\varkappa\zeta}{8} - \frac{(\Delta t)^2}{8\left(\Delta x\right)^2}\zeta\lambda^2 + \frac{\Delta t}{2\Delta x}\lambda\right| + \right.$$

$$\left.\left|\frac{1}{2} - \frac{\varkappa\zeta}{8} + \frac{(\Delta t)^2}{8\left(\Delta x\right)^2}\zeta\lambda^2 - \frac{\Delta t}{2\Delta x}\lambda\right|\right) \leqslant 1. \tag{158}$$

By virtue of (158) we find that (156) will be the necessary condition for the variational S-monotonicity (see Definition 1) of the non-linear scheme (147).

The variational scheme corresponding to (148) (under $\mathbf{v}_i^n = \mathbf{C} = const$) reads

$$\delta\mathbf{v}_i^{n+0.75} = \delta\mathbf{v}_i^{n+1} - \frac{\gamma}{2}\mathbf{G}_c\cdot\delta\mathbf{v}_i^{n+1}, \quad \gamma \equiv \frac{\Delta t}{\tau},$$

$$\delta\mathbf{v}_i^{n+1} = \delta\mathbf{v}_i^{n+0.5} + \gamma\mathbf{G}_c\cdot\delta\mathbf{v}_i^{n+0.75}, \quad \mathbf{G}_c = \mathbf{G}\left(C\right), \tag{159}$$

where $\mathbf{G}\left(\mathbf{u}\right) = \partial\mathbf{q}\left(\mathbf{u}\right)/\partial\mathbf{u}$. Let $\mathbf{G}$ will be a normal matrix. In such a case, since all eigenvalues of $\mathbf{G}_c$ have non-positive real parts, see (3), the first step of the scheme (148) will be S-monotone (see [10, Theorem 3.3], Theorem 2). It remains to prove that the scheme (159), taken as a whole, will be S-monotone under the same condition. Eliminating $\delta\mathbf{v}_i^{n+0.75}$ we obtain that

$$\delta\mathbf{v}_i^{n+1} = \left[\mathbf{I} - \gamma\mathbf{G}_c\cdot\left(\mathbf{I} - \frac{\gamma}{2}\mathbf{G}_c\right)\right]^{-1}\cdot\delta\mathbf{v}_i^{n+0.5}. \tag{160}$$

In view of [10, Theorem 3.3], the scheme (160) will be S-monotone if

$$\max_k\left|\frac{1}{1 - \gamma\xi_k\left(\mathbf{C}\right)\left(1 - \frac{\gamma}{2}\xi_k\left(\mathbf{C}\right)\right)}\right| \leq 1, \quad \forall\mathbf{C}\in\Omega_{\mathbf{u}}, \tag{161}$$

where $\xi_k\left(\mathbf{C}\right)$ denotes the k-th eigenvalue of the matrix $\mathbf{G}_c$. Since all eigenvalues of $\mathbf{G}_c$ have non-positive real parts, the variational scheme (160) will be unconditionally S-monotone.

Thus, the COS2 scheme, (147)-(148), will be variationally S-monotone only if (156) will be valid.

4.3 Second order schemes based on operator splitting techniques

Using the first order schemes (128) and (129), it can be constructed (see Section 6, Proposition 12) a scheme approximating (1) with the accuracy $O((\Delta x)^2 + (\Delta t)^2)$. Actually, since

$$\frac{(\Delta t)^2}{32}\left(\frac{\partial^2 \mathbf{u}}{\partial t^2}\right)_{i+0.5}^{n+0.25} = \frac{(\Delta t)^2}{32}\left(\frac{\partial^2 \mathbf{u}}{\partial t^2}\right)_{i}^{n+0.25} + O\left(\Delta x\,(\Delta t)^2\right), \tag{162}$$

the scheme will be as follows

$$\mathbf{v}_i^{n+0.25} = \mathbf{v}_i^n + \frac{\Delta t}{2\tau}\mathbf{q}\left(\mathbf{v}_i^{n+0.25}\right), \tag{163}$$

$$\mathbf{v}_{i+0.5}^{n+0.5} = 0.5\left(\mathbf{v}_{i+1}^{n+0.25} + \mathbf{v}_i^{n+0.25}\right) - \varkappa\frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^{n+0.25} - \mathbf{d}_i^{n+0.25}\right) - \frac{\Delta t}{2}\frac{\mathbf{f}\left(\mathbf{v}_{i+1}^{n+0.25}\right) - \mathbf{f}\left(\mathbf{v}_i^{n+0.25}\right)}{\Delta x}. \tag{164}$$

Hence, the stability as well as monotonicity behavior of (164) and (163) can be interesting itself. As demonstrated in Section 5, the behavior of Scheme (164) is similar to the one of the NT scheme, i.e. it can produce spurious oscillations if CFL number is not sufficiently small. Let us develop another scheme approximating (1) with the accuracy $O((\Delta x)^2+(\Delta t)^2)$ and such that its components (after operator splitting) will be of the second order. It can be done on the basis of the second order scheme (163)-(164) with ease. Actually, adding to and subtracting from Equation (229) (see Section 6, Proposition 12) the same quantity (162), we obtain (after operator splitting) the following scheme, instead of (163)-(164),

$$\mathbf{v}_i^{n+0.25} = \mathbf{v}_i^n + \frac{\Delta t}{2\tau}\mathbf{q}_i^{n+0.25} - \frac{(\Delta t)^2}{32}\left(\frac{\partial^2 \mathbf{u}}{\partial t^2}\right)_{i}^{n+0.25}, \tag{165}$$

$$\mathbf{v}_{i+0.5}^{n+0.5} = 0.5\left(\mathbf{v}_{i+1}^{n+0.25} + \mathbf{v}_i^{n+0.25}\right) - \varkappa\frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^{n+0.25} - \mathbf{d}_i^{n+0.25}\right) + \frac{(\Delta t)^2}{32}\left(\frac{\partial^2 \mathbf{u}}{\partial t^2}\right)_{i+0.5}^{n+0.25} - \frac{\Delta t}{2}\frac{\mathbf{f}\left(\mathbf{v}_{i+1}^{n+0.25}\right) - \mathbf{f}\left(\mathbf{v}_i^{n+0.25}\right)}{\Delta x}. \tag{166}$$

Thus, Scheme (165) as well as Scheme (166) are of the second order, and Scheme (165)-(166), taken as a whole, is of the second order as well.

Using Taylor series expansion, and central differencing, we find

$$\mathbf{v}_i^{n+0.125} = \mathbf{v}_i^{n+0.25} - \frac{\Delta t}{8}\left(\frac{\partial \mathbf{u}}{\partial t}\right)_{i}^{n+0.25} + \frac{1}{2}\left(\frac{\Delta t}{8}\right)^2\left(\frac{\partial^2 \mathbf{u}}{\partial t^2}\right)_{i}^{n+0.25}, \tag{167}$$

$$\mathbf{v}_i^{n+0.25} = \mathbf{v}_i^n + \frac{\Delta t}{4}\left(\frac{\partial \mathbf{u}}{\partial t}\right)_{i}^{n+0.125} + O\left((\Delta t)^3\right). \tag{168}$$

Considering the equation in (108) at $t_{n+0.25} < t \leq t_{n+0.5}$, we obtain, in view of (146), that

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} = -2\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{f}}{\partial x}\right) = -2\frac{\partial}{\partial x}\left(\frac{\partial \mathbf{f}}{\partial t}\right) = 4\frac{\partial}{\partial x}\left(\mathbf{A}^2 \cdot \frac{\partial \mathbf{u}}{\partial x}\right), \tag{169}$$

where $\mathbf{A} = \partial \mathbf{f} / \partial \mathbf{u}$. Then

$$\left[\frac{\partial}{\partial x}\left(\mathbf{A}^2 \cdot \frac{\partial \mathbf{u}}{\partial x}\right)\right]_{i+0.5}^{n+0.25} =$$

$$\frac{1}{\Delta x}\left[\left(\mathbf{A}_{i+1}^{n+0.25}\right)^2 \cdot \mathbf{d}_{i+1}^{n} - \left(\mathbf{A}_{i}^{n+0.25}\right)^2 \cdot \mathbf{d}_{i}^{n}\right] + O\left((\Delta x)^2\right). \tag{170}$$

By virtue of (109), (167)-(170), we rewrite Scheme (165)-(166) to read

$$\mathbf{v}_i^{n+0.125} = \mathbf{v}_i^{n+0.25} - \frac{\Delta t}{8\tau}\left(\mathbf{q}_i^{n+0.125} + \mathbf{q}_i^{n+0.25}\right),$$

$$\mathbf{v}_i^{n+0.25} = \mathbf{v}_i^{n} + \frac{\Delta t}{2\tau}\mathbf{q}_i^{n+0.125}, \tag{171}$$

$$\mathbf{v}_{i+0.5}^{n+0.5} = 0.5\left(\mathbf{v}_{i+1}^{n+0.25} + \mathbf{v}_{i}^{n+0.25}\right) - \varkappa\frac{\Delta x}{8}\left(\mathbf{d}_{i+1}^{n+0.25} - \mathbf{d}_{i}^{n+0.25}\right) +$$

$$\frac{(\Delta t)^2}{8\Delta x}\left[\left(\mathbf{A}_{i+1}^{n+0.25}\right)^2 \cdot \mathbf{d}_{i+1}^{n+0.25} - \left(\mathbf{A}_{i}^{n+0.25}\right)^2 \cdot \mathbf{d}_{i}^{n+0.25}\right]$$

$$- \frac{\Delta t}{2}\frac{\mathbf{f}\left(\mathbf{v}_{i+1}^{n+0.25}\right) - \mathbf{f}\left(\mathbf{v}_{i}^{n+0.25}\right)}{\Delta x}. \tag{172}$$

Notice, Scheme (172) coincides, in fact, with Scheme (147), however there is a difference between (171) and (148). In spite of the fact that both of these implicit Runge-Kutta schemes, (171) and (148), are of the second order, only the above combination, i.e. Scheme (171)-(172), is of the second order as a whole. Thus, even if a higher order implicit Runge-Kutta scheme will be used to approximate (109), nevertheless, there is a risk that Equation (1) will be approximated with the first order in time.

It can be proven, by analogy with the proof in Section 4.2, that (156) is the sufficient condition for stability as well as the necessary condition for S-monotonicity of Scheme (171)-(172).

There exists one more problem associated with the stiffness, i.e. $\tau \ll 1$, of (109). To demonstrate it, let us consider the following system of ordinary differential equations (ODEs) with a relaxation operator in the right-hand side:

$$\frac{dX(t)}{dt} = -\frac{\varphi(X,Y)}{\tau}(X - Y), \quad X(0) = X_0, \tag{173}$$

$$\frac{dY(t)}{dt} = -\frac{\varphi(X,Y)}{\tau}(Y - X), \quad Y(0) = Y_0. \tag{174}$$

The system will be considered at the interval $0 < t \leq \Delta t$. The following notation will be used: $X_\nu = X(\nu \Delta t)$. If $\varphi(X, Y) \equiv 1$, then the approximation, X_1^*, to X_1, in the case when the first order scheme (129) is used, will be

$$X_1^* = Z_0 + \frac{X_0 - Z_0}{1 + \lambda}, \tag{175}$$

where $Z_0 = 0.5(X_0 + Y_0)$, $\lambda = 2\Delta t / \tau$. In the case of the second order scheme, (148), the approximation to X_1 will be

$$X_1^{**} = Z_0 + \frac{X_0 - Z_0}{1 + \lambda + 0.5\lambda^2}. \tag{176}$$

The analytic solution, X_1, will be the following

$$X_1 = Z_0 + \frac{X_0 - Z_0}{\exp(\lambda)}. \tag{177}$$

Notice, in the case of underresolved numerical scheme we obtain that $\lambda \gg 1$. A comparison between (177) and (176) shows that there is a need to develop an implicit Runge-Kutta scheme of higher order accuracy, and such that the scheme (based on the operator splitting techniques), taken as a whole, will be, at least, of the second order. However, the higher order will be the implicit Runge-Kutta scheme, the more time-consuming procedure will be obtained. To tide over this problem, a semi-analytic approach could be useful. Let us demonstrate one, as applied to System (173)-(174). Let $\theta_\nu \equiv \varphi(X_\nu, Y_\nu)$. Integrating (173)-(174), and using the midpoint rule, we obtain

$$X_1 - X_0 = -\frac{\theta_{0.5}}{\tau} \int_0^{\Delta t} (X - Y)\, dt + O\left((\Delta t)^2\right), \tag{178}$$

$$Y_1 - Y_0 = -\frac{\theta_{0.5}}{\tau} \int_0^{\Delta t} (Y - X)\, dt + O\left((\Delta t)^2\right). \tag{179}$$

To resolve (178)-(179), a linearized version of System (173)-(174) can be used, namely, instead of $\varphi(X, Y)$ in (173)-(174), it can be taken $\theta_{0.5} = const$. The analytic solution will be the following

$$X_1 = Z_0 + \frac{X_0 - Z_0}{\exp(2\theta_{0.5}\Delta t / \tau)}, \quad Y_1 = Z_0 + \frac{Y_0 - Z_0}{\exp(2\theta_{0.5}\Delta t / \tau)}. \tag{180}$$

It remains to estimate the value of $\theta_{0.5}$. Since $\theta_{0.5} = 0.5(\varphi(X_0, Y_0) + \varphi(X_1, Y_1)) + O((\Delta t)^2)$, we obtain the following, in general, non-linear equation in $\theta_{0.5}$

$$\theta_{0.5} = 0.5(\varphi(X_0, Y_0) + \varphi(X_1, Y_1)), \tag{181}$$

where X_1 and Y_1 are defined via (180).

5 Exemplification and discussion

In this section we are mainly concerned with verification of the second order central scheme, (147). To demonstrate the superiority of this scheme over Scheme (128) being $O((\Delta x)^2 + \Delta t)$ accurate, we will use the inviscid Burgers equation. The COS2 scheme, (147)-(148), is verified using Pember's rarefaction test problem [55]. Euler equations of gas dynamics, namely Sod's as well as Lax problems, are also used for the verification of the second order central scheme, (147).

5.1 Scalar non-linear equation

As the first stage in the verification, we will focus on the following scalar version of the problem (1):

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \quad x \in \mathbb{R},\ 0 < t \le T_{\max}; \quad u(x,t)|_{t=0} = u^0(x). \tag{182}$$

To test the first order scheme, (128), we will solve the inviscid Burgers equation (i.e. $f(u) \equiv u^2/2$) with the following initial condition

$$u(x,0) = \begin{cases} u_0, & x \in (h_L, h_R) \\ 0, & x \notin (h_L, h_R) \end{cases}, \quad h_R > h_L,\ u_0 = const \neq 0. \tag{183}$$

The exact solution to (182), (183) is given by

$$u(x,t) = \begin{cases} u_1(x,t), & 0 < t \le T \\ u_2(x,t), & t > T \end{cases}, \tag{184}$$

where $T = 2S/u_0$, $S = h_R - h_L$,

$$u_1(x,t) = \begin{cases} \frac{x-h_L}{b-h_L} u_0, & h_L < x \le b,\ b = u_0 t + h_L \\ u_0, & b < x \le 0.5u_0 t + h_R \\ 0, & x \le h_L \ or \ x > 0.5u_0 t + h_R \end{cases}, \tag{185}$$

$$u_2(x,t) = \begin{cases} \frac{2S(x-h_L)}{(L-h_L)^2} u_0, & h_L < x \le L \\ 0, & x \le h_L \ or \ x > L \end{cases}, \tag{186}$$

$$L = 2\sqrt{S^2 + 0.5u_0 S(t-T)} + h_L. \tag{187}$$

The numerical solutions were computed on a uniform grid with spatial increments of $\Delta x = 0.01$, the velocity $u_0 = 1$ in (183), $h_L = 0.2$, $h_R = 1$, the monotonicity parameter $\aleph = 0.5$, the Courant number $Cr \equiv u_0 \Delta t/\Delta x = 0.5$, and the parameter $\varkappa = 1$ in (128). The results of simulation are depicted with the exact solution in Figure 2. We note (Figure 2) that the first order scheme, (128), exhibits a typical second-order nature, however spurious solutions are produced by the scheme. To obtain necessary conditions for the variational GOS-monotonicity (see Definition 1) of the COS1 scheme, (128), we consider

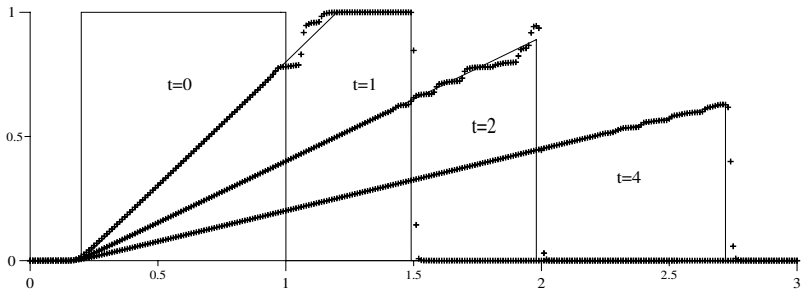

Figure 2: Inviscid Burgers equation. The scheme COS1 ($\varkappa = 1$) versus the analytical solution. Crosses: numerical solution; Solid line: analytical solution and initial data. $C_r = \aleph = 0.5$, $\Delta x = 0.01$.

the scheme in variations (132), corresponding to (128), on a constant ($v_i^n = C = const$, $\forall i, n$), i.e. (140). By virtue of Theorem 8 and Theorem 5, we find that, in the case of the Burgers equation, Scheme (140) will be GOS-monotone only if (139) will be fulfilled. Since all of the coefficients of (140) will be non-negative (the scalar version is considered) under (139), the scheme in question, (128), will be H-monotone (and hence TVD and G-monotone) only if (139) will be valid. Notice, the numerical simulations were performed with such values of the parameter $\varkappa$, Courant number, Cr, and monotonicity parameter, $\aleph$, that (139) was not violated. As it can be seen in Figure 2, the boundary maximum principle is not violated by the scheme, i.e., the maximum positive values of the dependent variable, v, occur at the boundary $t = 0$. It is interesting that the spurious solution (see Figure 2) produced by the scheme COS1 has the monotonicity property [26], since no new local extrema in x are created as well as the value of a local minimum is non-decreasing and the value of a local maximum is non-increasing.

Let us note that the problem of building free-of-spurious-oscillations schemes is, in general, unsettled up to the present. Even the best modern high-resolution schemes can produce spurious oscillations, and these oscillations are often of ENO type (see, e.g., [53] and references therein). We found that the oscillations produced by the COS1 scheme, (128), are of ENO type, namely their amplitude decreases rapidly with decreasing the time-increment Δt, and the oscillations virtually disappear under a relatively low Courant number, $Cr \leq 0.15$. However, the reduction of the Courant number causes some smearing of the solution. The spurious oscillations (see Figure 2) can be eradicated without reduction Courant number, but decreasing the parameter $\varkappa$. Particularly, the spurious oscillations disappear if $\varkappa = 2/3$, $C_r = 0.5$, however, this introduces more numerical smearing than in the case of the Courant number reduction. Satisfactory results are obtained under $\varkappa = 0.82$ ($C_r = 0.5$). The results of simulations are not depicted here.

To gain insight to why the scheme COS1, (128), can exhibit spurious solutions, let us consider the, so called, *first differential approximation* of this

scheme ([17, p. 45], [60, p. 376]; see also 'modified equations' in [17, p. 45], [40], [44]). As reported in [17], [60], this heuristic method was originally presented by Hirt (1968) (see [17, p. 45]) as well as by Shokin and Yanenko (1968) (see [60, p. 376]), and has since been widely employed in the development of stable difference schemes for PDEs.

We found that the local truncation error, ψ, for the scheme COS1 can be written in the following form

$$\psi = \frac{(1-\varkappa)(\Delta x)^2}{4\Delta t}\frac{\partial^2 u(x,t)}{\partial x^2} + \frac{\Delta t}{4}\frac{\partial^2 f(u)}{\partial t \partial x} + O\left(\frac{(\Delta x)^4}{\Delta t} + (\Delta t)^2 + (\Delta x)^2\right). \tag{188}$$

By virtue of (188), we find the first differential approximation of the scheme COS1

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \frac{\Delta t}{4}\frac{\partial}{\partial x}\left(B\frac{\partial u(x,t)}{\partial x}\right), \tag{189}$$

where $B = (1-\varkappa)(\Delta x / \Delta t)^2 - A^2$. The term in right-hand side of (189) will be dissipative if

$$(1-\varkappa)\left(\frac{\Delta x}{\Delta t}\right)^2 - A^2 > 0, \implies C_r^2 < 1-\varkappa. \tag{190}$$

Thus, the scheme COS1, (128), is non-dissipative under $\varkappa = 1$, and hence can produce spurious oscillations. Notice, if $\varkappa = 0.82$, then we obtain from (190) that $C_r < 0.42$. Nevertheless, as it is reported above, satisfactory results can be obtained under $C_r = 0.5$ as well.

So then, the notion of first differential approximation has enabled us to understand that the spurious solutions exhibited by the scheme COS1, (128), are mainly associated with the negative numerical viscosity introduced to obtain the scheme of the second order in space, i.e. $O((\Delta x)^2 + \Delta t)$. Let us consider the scheme COS2, (147), approximating (182) with the accuracy $O((\Delta x)^2 + (\Delta t)^2)$. Notice, the second order scheme COS2, (147), is nothing more than the scheme COS1, (128), with the additional non-negative numerical viscosity. To test the scheme COS2, (147), the inviscid Burgers equation was solved under the initial condition (183). The numerical solutions were computed under the same values of parameters as in the case of the scheme COS1, but $C_r = 1$. The results of simulation are depicted with the exact solution in Figure 3.

We note (Figure 3) that the scheme COS2, (147), exhibits a typical second-order nature without any spurious oscillations. Increasing the value of $\aleph$ (up to $\aleph = 1$) leads to a minor improvement of the numerical solutions, whereas decreasing the value of Cr leads to a mild smearing of the solutions. The results of simulations are not depicted here.

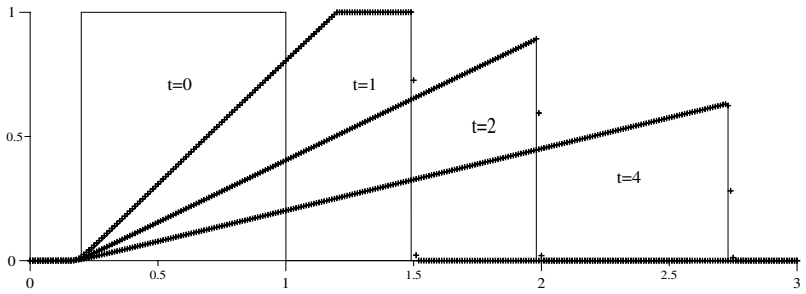

Figure 3: Inviscid Burgers equation. The scheme COS2 ($\varkappa = 1$) versus the analytical solution. Crosses: numerical solution; Solid line: analytical solution and initial data. $C_r = 1$, $\aleph = 0.5$, $\Delta x = 0.01$.

5.2 Hyperbolic conservation laws with relaxation

Let us consider the model system of hyperbolic conservation laws with relaxation developed in [55]:

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u^2 + aw\right) = 0, \tag{191}$$

$$\frac{\partial z}{\partial t} + \frac{\partial}{\partial x}az = \frac{1}{\tau}Q(w,z), \tag{192}$$

where

$$Q(w,z) = z - m(u - u_0), \quad u = w - q_0 z, \tag{193}$$

τ denotes the relaxation time of the system, q_0, m, a, and u_0 are constants. The Jacobian, $\mathbf{A}$, can be written in the form

$$\mathbf{A} = \left\{ \begin{array}{cc} w - q_0 z + a & -q_0\left(w - q_0 z\right) \\ 0 & a \end{array} \right\}. \tag{194}$$

The system (191)-(192) has the following frozen [55] characteristic speeds $\lambda_1 = a$, $\lambda_2 = u + a$. The equilibrium equation for (191)-(192) is

$$\frac{\partial w}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}u_*^2 + aw\right) = 0, \tag{195}$$

where

$$u_* = w - q_0 z_*, \quad z_* = \frac{m}{1 + mq_0}\left(w - u_0\right). \tag{196}$$

The equilibrium characteristic speed λ_* can be written in the form

$$\lambda_*\left(w\right) = \frac{u_*\left(w\right)}{1 + mq_0} + a. \tag{197}$$

Pember's rarefaction test problem is to find the solution $\{w, z\}$ to (191)-(192), and hence the function $u = u\left(x,t\right)$, under $\tau \to 0$, and where

$$\{w,z\} = \left\{ \begin{array}{ll} \left\{w_L, z_*\left(w_L\right)\right\}, & x < x_0 \\ \left\{w_R, z_*\left(w_R\right)\right\}, & x > x_0 \end{array} \right. , \tag{198}$$

$$0 < u_L = w_L - q_0 z_* (w_L) < u_R = w_R - q_0 z_* (w_R) . \tag{199}$$

The analytical solution of this problem can be found in [55]. The parameters of the model system are assumed as follows: $q_0 = -1$, $m = -1$, $u_0 = 3$, $a = \pm 1$, $\tau = 10^{-8}$. The initial conditions of the rarefaction problem are defined by

$$u_L = 2, \implies z_L = m (u_L - u_0) = 1, \; w_L = u_L + q_0 z_L = 1, \tag{200}$$

$$u_R = 3, \implies z_R = m (u_R - u_0) = 0, \; w_R = u_R + q_0 z_R = 3. \tag{201}$$

The position of the initial discontinuity, x_0, is set according to the value of a so that the solutions of all the rarefaction problems are identical [55]. Let a position, x_R^t, of leading edge or a position, x_L^t, of trailing edge of the rarefaction be known (e.g., $x_R^t = 0.85$, $x_L^t = 0.7$ in [55]), then

$$x_0 = x_R^t - \left(\frac{u_R}{1 + m q_0} + a \right) t = x_L^t - \left(\frac{u_L}{1 + m q_0} + a \right) t. \tag{202}$$

At $t = 0.3$, under (200)-(201) we have [55]

$$u = \begin{cases} 2, & x \le 0.7 \\ 2 + \frac{x - 0.7}{0.85 - 0.7}, & 0.7 < x < 0.85 \\ 3, & x \ge 0.85 \end{cases} . \tag{203}$$

The results of simulations, based upon the scheme COS2, (147)-(148), under different values of the parameter a ($a = 1$, $a = -1$) and different values of a grid spacing, Δx, are depicted in Figure 4.

One can clearly see (Figure 4) that the scheme COS2 is free from spurious oscillations. Let us also note that the results generated by the scheme COS2 are less accurate in the case of negative value of a than those in the case of positive value of a. Specifically, in the numerical solutions produced under $a = -1$, the representations of the trailing and leading edges of the rarefaction are more smeared than those in the solutions produced under $a = 1$. Notice, under some negative value of a, the frozen and the equilibrium characteristic speeds do not all have the same sign.

5.3 Euler equations of gas dynamics

In this subsection we apply the second order scheme COS2, (147), to the Euler equations of gamma-law gas:

$$\frac{\partial \mathbf{u} (x,t)}{\partial t} + \frac{\partial}{\partial x} \mathbf{F} (\mathbf{u}) = 0, \quad x \in \mathbb{R}, \; t > 0; \quad \mathbf{u} (x,0) = \mathbf{u}^0 (x) , \tag{204}$$

$$\mathbf{u} \equiv \{u_1, u_2, u_3\}^T = \{\rho, \rho v, e\}^T , \quad \mathbf{F} (\mathbf{u}) = \left\{ \rho v, \rho v^2 + p, (e + p) v \right\}^T , \tag{205}$$

$$e = \frac{p}{\gamma - 1} + \frac{1}{2} \rho v^2, \quad \gamma = const, \tag{206}$$

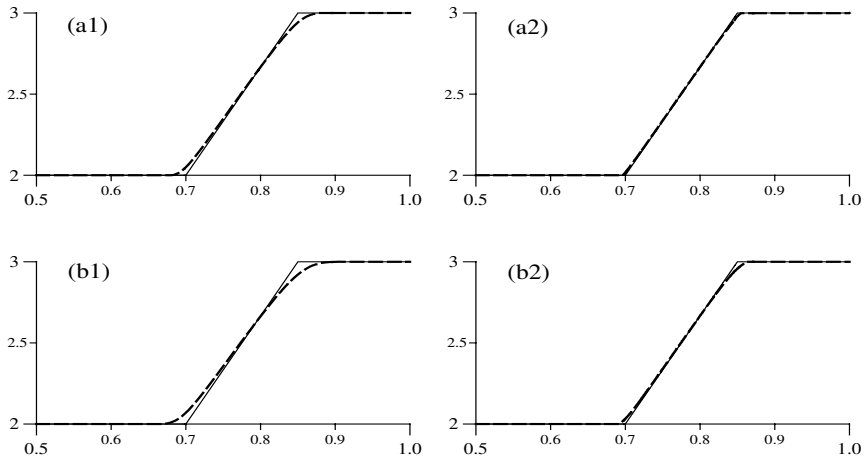

Figure 4: Pember's rarefaction test problem. The second-order scheme COS2 ($\varkappa = 1$) versus the analytical solution for u. Dashed line: numerical solution; Solid line: analytical solution. Time $t = 0.3$, Courant number $C_r = 1$, monotonicity parameter $\aleph = 1$. (a1): $\Delta x = 10^{-3}$, $a = 1$; (a2): $\Delta x = 2.5 \times 10^{-4}$, $a = 1$; (b1): $\Delta x = 10^{-3}$, $a = -1$; (b2): $\Delta x = 2.5 \times 10^{-4}$, $a = -1$.

where ρ, v, p, e denote the density, velocity, pressure, and total energy respectively. We consider the Riemann problem subject to Riemann initial data

$$\mathbf{u}^0(x) = \begin{cases} \mathbf{u}_L & x < x_0 \\ \mathbf{u}_R & x > x_0 \end{cases}, \quad \mathbf{u}_L, \mathbf{u}_R = const. \tag{207}$$

The analytic solution to the Riemann problem can be found in [40, Sec. 14].

Aiming to suppress possible spurious oscillations, the scheme COS2, (147), is modified in such a way as to prevent a violation of the necessary conditions (see Theorem 5) for variational monotonicity of the scheme COS2, (147). The modification is to use $\aleph = \aleph(x,t)$ instead of $\aleph = const$. Leaving out the terms equal to zero under $\mathbf{v}_i^n = \mathbf{C} = const$ in (147), we write the reduced variational scheme corresponding to the scheme COS2 in the following form

$$\delta\mathbf{v}_{i+0.5}^{n+0.5} = \mathbf{C}_i^n \cdot \delta\mathbf{v}_i^n + \mathbf{D}_i^n \cdot \delta\mathbf{v}_{i+1}^n, \tag{208}$$

$$\delta\mathbf{v}_i^{n+1} = \mathbf{C}_{i-0.5}^{n+0.5} \cdot \delta\mathbf{v}_{i-0.5}^{n+0.5} + \mathbf{D}_{i-0.5}^{n+0.5} \cdot \delta\mathbf{v}_{i+0.5}^{n+0.5}, \tag{209}$$

where

$$\mathbf{C}_i^n = \frac{1}{2}\left(\mathbf{I} + \frac{\Delta t}{\Delta x}\mathbf{A}_i^n + \mathbf{B}_i^n\right), \; \mathbf{D}_i^n = \frac{1}{2}\left(\mathbf{I} - \frac{\Delta t}{\Delta x}\mathbf{A}_{i+1}^n - \mathbf{B}_i^n\right), \tag{210}$$

$$\mathbf{B}_i^n = \frac{1}{4}\left(\mathbb{B}_i^n - \mathbb{A}_i^n\right) - \frac{\Delta t^2}{4\Delta x^2}\left[\left(\mathbf{A}_{i+1}^n\right)^2 \cdot \mathbb{B}_i^n - \left(\mathbf{A}_i^n\right)^2 \cdot \mathbb{A}_i^n\right], \tag{211}$$

$$\mathbf{C}_{i-0.5}^{n+0.5} = \frac{1}{2}\left(\mathbf{I} + \frac{\Delta t}{\Delta x}\mathbf{A}_{i-0.5}^{n+0.5} + \mathbf{B}_{i-0.5}^{n+0.5}\right), \tag{212}$$

$$\mathbf{D}_{i-0.5}^{n+0.5} = \frac{1}{2}\left(\mathbf{I} - \frac{\Delta t}{\Delta x}\mathbf{A}_{i+0.5}^{n+0.5} - \mathbf{B}_{i-0.5}^{n+0.5}\right), \tag{213}$$

$$\mathbf{B}_{i-0.5}^{n+0.5} = \frac{1}{4}\left(\mathbb{B}_{i-0.5}^{n+0.5} - \mathbb{A}_{i-0.5}^{n+0.5}\right) -$$

$$\frac{\Delta t^2}{4\Delta x^2}\left[\left(\mathbf{A}_{i+0.5}^{n+0.5}\right)^2 \cdot \mathbb{B}_{i-0.5}^{n+0.5} - \left(\mathbf{A}_{i-0.5}^{n+0.5}\right)^2 \cdot \mathbb{A}_{i-0.5}^{n+0.5}\right]. \tag{214}$$

Notice, if the diagonal elements of the matrices $\mathbf{C}_i^n$ and $\mathbf{D}_i^n$ are non-negative, i.e. $C_i^{n,kk} \geq 0$ and $D_i^{n,kk} \geq 0$, then any matrix column of the scheme (208) consists of zeros, but two entries. In such a case this column is a μ-function. Hence, at the first stage of modification, it is sufficient to verify $C_i^{n,kk}$ and $D_i^{n,kk}$. If $C_i^{n,kk} < 0$ or $D_i^{n,kk} < 0$ for, at least, one value of k, then the value of $\aleph$ at the node i (and, if it is necessary, at the neighbor nodes) is reduced up to $\aleph_{\min} \geq 0$. Then, the modified values of $\aleph$, i.e. $\aleph_i^n$, are used in the scheme COS2, (147). After that we turn to the second stage of modifications. Eliminating $\delta\mathbf{v}_{i+0.5}^{n+0.5}$ from (208)-(209) we convert the variational scheme based on staggered spatial grid into the following non-staggered form

$$\delta\mathbf{v}_i^{n+1} = \mathbf{F}_i^n \cdot \delta\mathbf{v}_{i-1}^n + \mathbf{G}_i^n \cdot \delta\mathbf{v}_i^n + \mathbf{H}_i^n \cdot \delta\mathbf{v}_{i+1}^n, \tag{215}$$

where

$$\mathbf{F}_i^n = \mathbf{C}_{i-0.5}^{n+0.5} \cdot \mathbf{C}_{i-1}^n,\ \mathbf{G}_i^n = \mathbf{C}_{i-0.5}^{n+0.5} \cdot \mathbf{D}_{i-1}^n + \mathbf{D}_{i-0.5}^{n+0.5} \cdot \mathbf{C}_i^n,\ \mathbf{H}_i^n = \mathbf{D}_{i-0.5}^{n+0.5} \cdot \mathbf{D}_i^n. \tag{216}$$

In view of Theorem 5, we obtain that the following inequalities must be valid

$$F_i^{n,kk},\ G_i^{n,kk},\ H_i^{n,kk} \geq 0, \quad G_i^{n,kk} \geq \min\left\{F_{i+1}^{n,kk}, H_{i-1}^{n,kk}\right\} \quad \forall i,\ \forall k. \tag{217}$$

where $F_i^{n,kk}$, $G_i^{n,kk}$, $H_i^{n,kk}$ denote the diagonal elements of the matrices $\mathbf{F}_i^n$, $\mathbf{G}_i^n$, and $\mathbf{H}_i^n$, respectively. By analogy with the first stage, the value of $\aleph_i^n$ is modified at every node, where at least one inequality in (217) is violated.

First we solve the shock tube problem (see, e.g., [6], [40], [41]) with Sod's initial data:

$$\mathbf{u}_L = \left\{\begin{array}{c} 1 \\ 0 \\ 2.5 \end{array}\right\}, \quad \mathbf{u}_R = \left\{\begin{array}{c} 0.125 \\ 0 \\ 0.25 \end{array}\right\}. \tag{218}$$

Following Balaguer and Conde [6] as well as Liu and Tadmor [41] we assume that the computational domain is $0 \leq x \leq 1$; the point x_0 is located at the middle of the interval $[0, 1]$, i.e. $x_0 = 0.5$; the equations (204) are integrated up to $t = 0.16$ on a spatial grid with 200 nodes as in [6] and in [41]. The Courant number is taken to be $Cr = 1$ (or $Cr = 0.9$) in contrast to [6] and [41], where the simulations were done under $\Delta t = 0.1\Delta x$ (i.e. $0.13 \lesssim Cr \lesssim 0.22$). The results of simulations with the two-stage modification of the monotonicity parameter $\aleph$ are depicted in Figure 5.

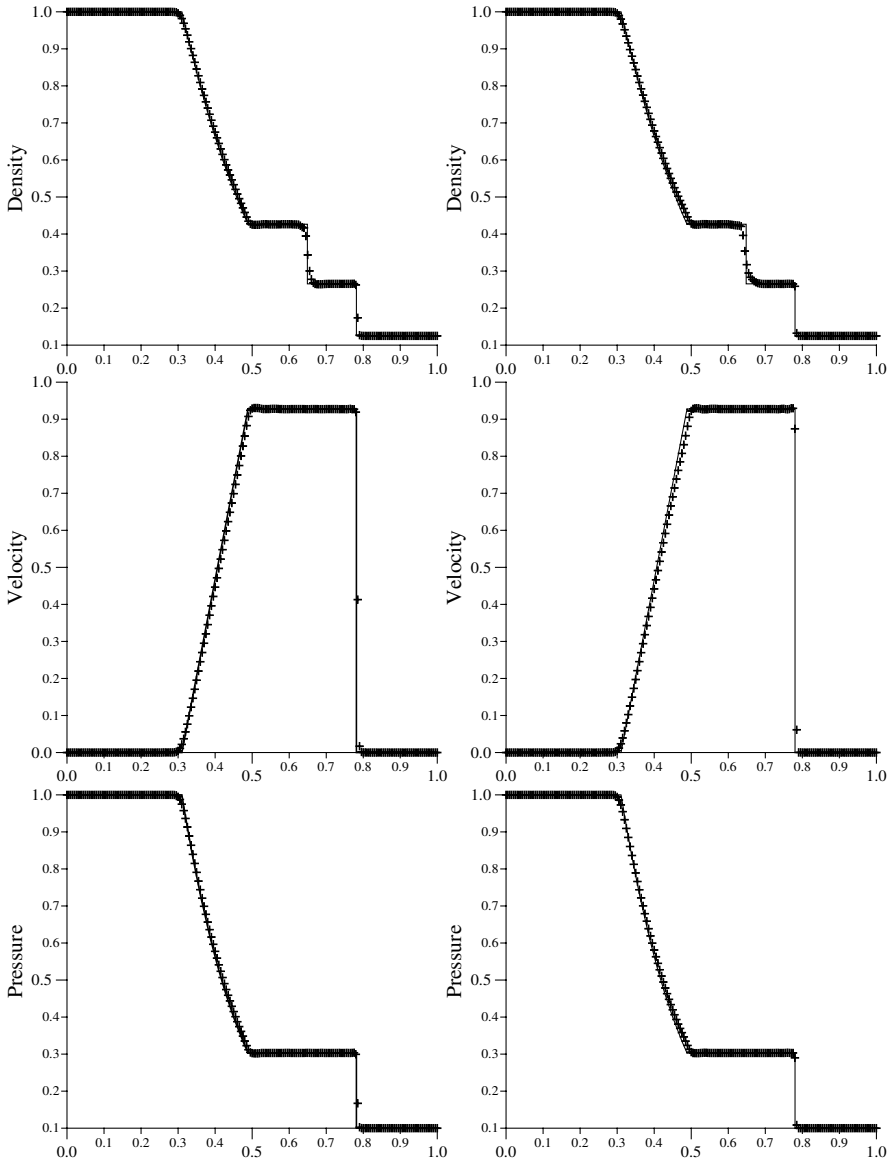

Figure 5: Sod's problem. The scheme COS2 with two-stage modification under $C_r = 0.9$, $\aleph \in [0.3, 1]$ (left column) and $C_r = 1$, $\aleph \in [0, 1]$ (right column) versus the analytical solution. Time $t = 0.16$, spatial increment $\Delta x = 0.005$.

The results depicted in Figure 5 are not worse in comparison to the corresponding third-order central results of [41, p. 418] as well as to the results obtained by the fourth-order non-oscillatory scheme in [6, p. 472]. Moreover, the scheme COS2, (147), does not produce any spurious oscillations. We continue the testing of the modified scheme COS2, (147), using the shock tube problem with Lax initial data

$$\mathbf{u}_L = \left\{ \begin{array}{c} 0.445 \\ 0.311 \\ 8.928 \end{array} \right\}, \quad \mathbf{u}_R = \left\{ \begin{array}{c} 0.5 \\ 0 \\ 1.4275 \end{array} \right\}. \tag{219}$$

The results of simulations with the two-stage modification under $C_r = 0.9$, $\aleph_{\min} = 0.3$ (left column) and $\aleph_{\min} = 0.001$ (right column) are depicted in Figure 6.

One can readily see (Figure 6) that the scheme COS2 gives a not worse resolution than that obtained by the schemes studied in [41] and [6, p. 472] without, in fact, spurious oscillations.

Thus, the algorithm based on Theorem 5 shows a possibility to avoid spurious numerical oscillations in a computed solution totally, and hence the second order scheme, COS2, is found to be accurate and robust. However this algorithm can lead to a smearing effect that partly reduces accuracy. There are several reasons for this side effect of the algorithm. Particularly, the algorithm is not optimal, i.e. it gives no more than rough approximations from below to the values of $\aleph$. Moreover, in this algorithm, the value of $\aleph$ at a grid node i is considered as a scalar, i.e. it is assumed that the monotonicity parameter, $\aleph$, is common to all coordinates of the vector $\mathbf{v}_i$. Thus, the algorithm lacks flexibility, and hence the accuracy of the scheme COS2, (147), can be reduced. Notice, the conditions of Theorem 5 are not necessary and sufficient conditions for monotonicity (i.e., absence of spurious oscillations) of the difference scheme COS2, and hence the total elimination of spurious oscillations on the basis of Theorem 5 can lead to an undesirable smearing effect. All this must be given proper weight in designing of the difference schemes.

6 Appendix

Proposition 11 *Let us find the order of accuracy, r, in (118) if d_i will be approximated by $\widetilde{d}_i$ with the order of accuracy s, i.e. let*

$$d_i = \widetilde{d}_i + O\left((\Delta x)^s\right). \tag{220}$$

Let $U(x)$ be sufficiently smooth, then we can write

$$U_{i+1} = U_{i+05} + U'_{i+05}\frac{\Delta x}{2} + \frac{1}{2}U''_{i+05}\left(\frac{\Delta x}{2}\right)^2 + O\left((\Delta x)^3\right), \tag{221}$$

$$U_i = U_{i+05} - U'_{i+05}\frac{\Delta x}{2} + \frac{1}{2}U''_{i+05}\left(\frac{\Delta x}{2}\right)^2 + O\left((\Delta x)^3\right). \tag{222}$$

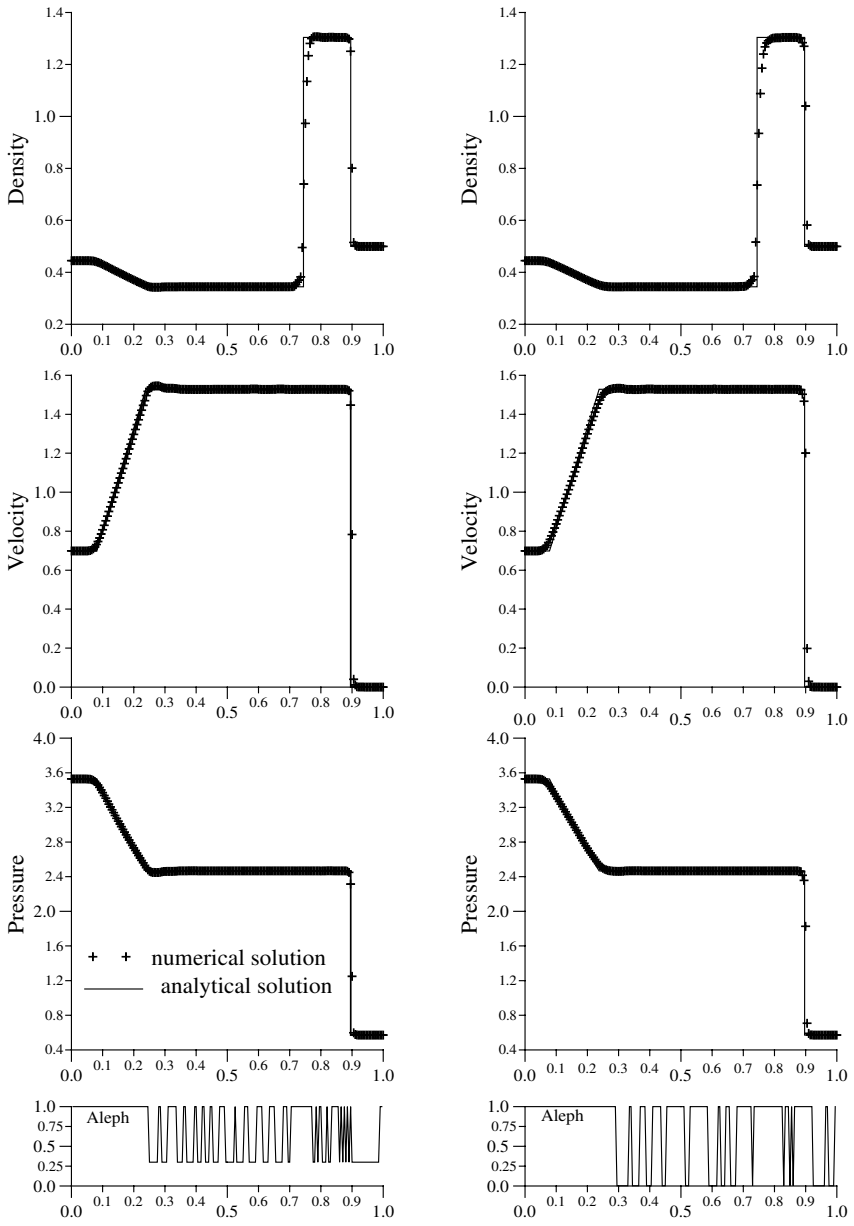

Figure 6: Lax problem. The scheme COS2 with two-stage modification under $C_r = 0.9$, $\aleph \in [0.3, 1]$ (left column) and $\aleph \in [0.001, 1]$ (right column) versus the analytical solution. Time $t = 0.16$, spatial increment $\Delta x = 0.005$, Aleph: the distribution of $\aleph$ after the second stage of modifications

Combining the equalities (221) and 222 we obtain

$$U_{i+1} + U_i = 2U_{i+05} + \left.\frac{\partial^2 U}{\partial x^2}\right|_{i+05} \left(\frac{\Delta x}{2}\right)^2 + O\left((\Delta x)^3\right). \tag{223}$$

In a similar manner we write:

$$d_{i+1} = U'_{i+05} + U''_{i+05}\frac{\Delta x}{2} + \frac{1}{2}U'''_{i+05}\left(\frac{\Delta x}{2}\right)^2 + O\left((\Delta x)^3\right), \tag{224}$$

$$d_i = U'_{i+05} - U''_{i+05}\frac{\Delta x}{2} + \frac{1}{2}U'''_{i+05}\left(\frac{\Delta x}{2}\right)^2 + O\left((\Delta x)^3\right). \tag{225}$$

Subtracting the equations (224) and (225), we obtain

$$\left.\frac{\partial^2 U}{\partial x^2}\right|_{i+05} = \frac{d_{i+1} - d_i}{\Delta x} + O\left((\Delta x)^2\right). \tag{226}$$

In view of (226) and (220) we obtain from (223) the following interpolation formula

$$U_{i+05} = \frac{1}{2}\left(U_{i+1} + U_i\right) - \frac{\Delta x}{8}\left(\widetilde{d}_{i+1} - \widetilde{d}_i\right) + O\left((\Delta x)^4 + (\Delta x)^{s+1}\right). \tag{227}$$

In view of (227) we obtain that $r = \min(4, s+1)$.

Proposition 12 *As applied to central schemes (see Section 4.1), we will construct a second order scheme based on operator-splitting techniques. We will, in fact, use the summarized (summed) approximation method [59, Section 9.3] to estimate order of approximation. Consider the following equation*

$$\mathcal{P}\mathbf{u} \equiv \mathcal{P}_1\mathbf{u} + \mathcal{P}_2\mathbf{u} \equiv \frac{\partial \mathbf{u}}{\partial t} - L\mathbf{u} = 0, \quad \mathcal{P}_k\mathbf{u} \equiv \frac{1}{2}\frac{\partial \mathbf{u}}{\partial t} - L_k\mathbf{u},\ k = 1, 2, \tag{228}$$

where L_k *is an operator, e.g. a differential operator, a real analytic function, etc', acting on* $\mathbf{u}(x,t)$. *We approximate (228) on the cell* $[x_i, x_{i+1}] \times [t_n, t_{n+0.5}]$ *by the following difference equation with the accuracy* $O((\Delta x)^2 + (\Delta t)^2)$

$$\Pi\mathbf{v} \equiv \frac{\mathbf{v}^{n+0.5}_{i+0.5} - \mathbf{v}^{n}_{i+0.5}}{0.5\Delta t} - \Lambda_1\mathbf{v}^{n+0.25} - \Lambda_2\mathbf{v}^{n+0.25} = 0, \tag{229}$$

where it is assumed that the operator $L_k\mathbf{u}$ *is approximated by the operator* $\Lambda_k\mathbf{u}$ *with the accuracy* $O((\Delta x)^2)$, *i.e.*

$$\Lambda_k\mathbf{u}^{n+0.25} = (L_k\mathbf{u})^{n+0.25}_{i+0.5} + O\left((\Delta x)^2\right). \tag{230}$$

In view of the operator splitting idea, to the problem (229) there corresponds the following chain of difference schemes

$$\Pi_1\mathbf{w} \equiv \frac{1}{2}\frac{\mathbf{w}^{n+0.25}_{i+0.5} - \mathbf{w}^{n}_{i+0.5}}{0.25\Delta t} - \Lambda_1\mathbf{w}^{n+0.25}_1 = 0, \tag{231}$$

$$\Pi_2 \mathbf{w} \equiv \frac{1}{2} \frac{\mathbf{w}_{i+0.5}^{n+0.5} - \mathbf{w}_{i+0.5}^{n+0.25}}{0.25\Delta t} - \Lambda_2 \mathbf{w}_2^{n+0.25} = 0. \tag{232}$$

One can see from the above that the operator $\mathcal{P}_k \mathbf{u}$ *is approximated by* $\Pi_k \mathbf{u}$ *with the accuracy* $O(\Delta t + (\Delta x)^2)$

$$\Pi_1 \mathbf{u}_{i+0.5}^{n+0.25} = (\mathcal{P}_1 \mathbf{u})_{i+0.5}^{n+0.25} - \frac{\Delta t}{16} \left(\frac{\partial^2 \mathbf{u}}{\partial t^2} \right)_{i+0.5}^{n+0.25} + O\left((\Delta t)^2 + (\Delta x)^2 \right), \tag{233}$$

$$\Pi_2 \mathbf{u}_{i+0.5}^{n+0.25} = (\mathcal{P}_2 \mathbf{u})_{i+0.5}^{n+0.25} + \frac{\Delta t}{16} \left(\frac{\partial^2 \mathbf{u}}{\partial t^2} \right)_{i+0.5}^{n+0.25} + O\left((\Delta t)^2 + (\Delta x)^2 \right). \tag{234}$$

In view of (233)-(234), the local truncation error [40, p. 142], ψ*, on a sufficiently smooth solution* $\mathbf{u}(x,t)$ *to (228) is found to be*

$$\psi = \Pi \mathbf{u} = \Pi_1 \mathbf{u} + \Pi_2 \mathbf{u} = \tag{235}$$

$$(\mathcal{P}_1 \mathbf{u} + \mathcal{P}_2 \mathbf{u})_{i+0.5}^{n+0.25} + O\left((\Delta t)^2 + (\Delta x)^2 \right) = O\left((\Delta t)^2 + (\Delta x)^2 \right). \tag{236}$$

Thus, implicit Scheme (231) together with explicit Scheme (232) approximate (228) with the second order.

References

[1] R. Abgrall and P. L. Roe, High Order Fluctuation Schemes on Triangular Meshes, Journal of Scientific Computing, Vol. 19, Nos. 1-3 (2003), pp. 3-36.

[2] I. Ahmad., M. Berzins, MOL solvers for hyperbolic PDEs with source terms, Mathematics and Computers in Simulation 56 (2001) 115-125

[3] D. A. Anderson, J. C. Tannehill and R. H. Pletcher*, Computational Fluid Mechanics and Heat Transfer*, Hemisphere Publishing Corporation, New York, 1984.

[4] Alan L. Andrew, K.-W. Eric Chu, Peter Lancaster, Derivatives of eigenvalues and eigenvectors of matrix functions, SIAM J. Matrix Anal. Appl., Vol. 14, No. 4 (1993), pp. 903-926.

[5] Mark A. Aves, David F. Griffiths, and Desmond J. Higham, Runge-Kutta solutions of a hyperbolic conservation law with source term, SIAM J. Sci. Comput., Vol. 22, No. 1, pp. 20-38 (2000)

[6] Angel Balaguer and Carlos Conde, Fourth-order non-oscillatory upwind and central schemes for hyperbolic conservation laws, SIAM J. Numer. Anal. , Vol. 43, No. 2, pp. 455-473, (2005)

[7] François Bereux, Lionel Sainsaulieu, A roe-type Riemann solver for hyperbolic systems with relaxation based on time-dependent wave decomposition, Numer. Math. 77: 143-185 (1997).

[8] I. Boglaev, Monotone Iterates for Solving Nonlinear Monotone Difference Schemes, Computing 78 (2006), 17-30.

[9] V. S. Borisov and M Mond, On monotonicity, stability, and construction of central schemes for hyperbolic conservation laws with source terms, (2007), arXiv:0705.1109v3 [physics.comp-ph].

[10] V. S. Borisov and S. Sorek, - On monotonicity of difference schemes for computational physics, SIAM J. Sci. Comput., Vol. 25, No. 5 (2004), pp. 1557-1584.

[11] V. S. Borisov, On discrete maximum principles for linear equation systems and monotonicity of difference schemes, SIAM J. Matrix Anal. Appl., Vol. 24, No. 4 (2003), pp. 1110-1135.

[12] Russel E. Caflisch, Shi Jin, and Giovanni Russo, Uniformly accurate schemes for hyperbolic systems with relaxation, SIAM J. Numer. Anal., Vol. 34, No. 1 (1997), pp. 246-281.

[13] S. R. Chacravarthy, S. Osher, High resolution applications of the Osher upwind scheme for the Euler equations, AIAA paper #83-1943, Danver, Mass. (19843), pp. 363-372.

[14] Tao Du, Jing Shi, Zi-Niu Wu, Mixed analytical/numerical method for flow equations with a source term, Computers & Fluids 32 (2003) 659-690.

[15] Duboshin G. N., *Theoretical foundations of stability of motion*, Moscow University, Moscow, 1952 (in Russian).

[16] F.N. Fritsch and R.E. Carlson, Monotone piecewise cubic interpolation, SIAM J. Numer. Anal. 17, No. 2, 238-246 (1980).

[17] V. G. Ganzha and E. V. Vorozhtsov, *Numerical Solutions for Partial Differential Equations,* CRC Press, New York, 1996.

[18] V. G. Ganzha and E. V. Vorozhtsov, *Computer-aided analysis of difference schemes for partial differential equations*, John Wiley & Sons, New York, 1996.

[19] Gil' M. I., *Stability of finite and infinite dimensional systems*, Kluwer Academic Publishers, Boston, 1998.

[20] Gil' M. I., *Difference Equations in Normed Spaces, Stability and Oscillations*, Elsevier, Amsterdam, 2007.

[21] Edwige Godlewski and Pierre-Arnaud Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws,* Springer-Verlag, New York, 1996.

[22] S. K. Godunov, A finite difference method for the numerical computation and discontinuous solutions of fluid dynamics, Mat. Sb., 47 (1959), pp. 271-306 (in Russian).

[23] L. Gosse, A Well-Balanced Flux-Vector Splitting Scheme Designed for Hyperbolic Systems of Conservation Laws with Source Terms, Computers and Mathematics with Applications 39 (2000) 135-159

[24] B. Hanouzet and R. Natalini, Global Existence of Smooth Solutions for Partially Dissipative Hyperbolic Systems with a Convex Entropy, Arch. Rational Mech. Anal. 169 (2003) 89-117.

[25] A. Harten, J. M. Hyman, and P. D. Lax, with appendix by B. Keyfiz, On finite-difference approximations and entropy conditions for shocks, Comm. Pure Appl. Math., 29 (1976), pp. 297-322.

[26] A. Harten, High resolution schemes for hyperbolic conservation laws, J. Comput. Phys., V. 49 (1983), pp. 357-393.

[27] Ami Harten, On a class of high resolution total-variaton-stable finite difference schemes, SIAM J. Numer. Anal., Vol. 21, No. 1 (1984), pp. 1-23.

[28] Ami Harten, Uniformly High Order Accurate Essentially Non-oscillatory Schemes, Ill, J. Comput. Phys., V. 71 (1987), pp. 231-303.

[29] Róbert Horváth, On the monotonicity conservation in numerical solutions of the heat equation, Appl. Numer. Math., 42 (2002), pp. 189-199.

[30] Shi Jin, Runge-Kutta Methods for Hyperbolic Conservation Laws with Stiff Relaxation Terms, J. Comp. Phys. 122 (1995), 51-67.

[31] Shi Jin, Lorenzo Pareschi, and Gioseppe Toscani, Uniformly accurate diffusive schemes for multiscale transport equations, SIAM J. Numer. Anal. Vol. 38, No. 3 (2000), pp. 913-936.

[32] Shi Jin and C. David Levermore, Numerical Schemes for Hyperbolic Conservation Laws with Stiff Relaxation Terms, J. Comp. Phys. 126, 449-467 (1996).

[33] David Kahaner, Cleve Moler, and Stephen Nash, *Numerical methods and software*, Prentice-Hall, New Jersey, 1989.

[34] N. N. Kalitkin, *Numerical Methods*, Nauka, Moscow, 1978 (in Russian).

[35] L.M. Kocić and G.V. Milovanović, Shape Preserving Approximations by Polynomials and Splines, Computers Math. Applic. Vol. 33, No. 11, pp. 59-97, 1997.

[36] A. G. Kulikovskii , N. V. Pogorelov, A. Yu. Semenov, *Mathematical problems of numerical solution of hyperbolic systems*, Nauka, Moscow, 2001 (in Russian).

[37] Alexander Kurganov and Eitan Tadmor, New High-Resolution Central Schemes for Nonlinear Conservation Laws and Convection-Diffusion Equations, Journal of Computational Physics 160, 241-282 (2000)

[38] Alexander Kurganov and Doron Levy, A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations, SIAM J. Sci. Comput., Vol. 22, No. 4 (2000), pp. 1461-1488

[39] P. Lancaster, *Theory of Matrices*, Academic Press, New York, 1969.

[40] Randall J. LeVeque, *Finite volume methods for hyperbolic problems*, Cambridge University Press, Cambridge, 2002.

[41] Xu-Dong Liu and Eitan Tadmor, Third order nonoscillatory central scheme for hyperbolic conservation laws, Numer. Math. (1998) 79: 397-425.

[42] L. A. Monthé, A study of splitting scheme for hyperbolic conservation laws, Journal of Computational and Applied Mathematics 137 (2001) 1-12.

[43] L. Mirsky, *An Introduction to Linear Algebra*, Dover Publications, New York, 1990.

[44] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, 1996.

[45] K. W. Morton, Discretization of unsteady hyperbolic conservation laws, SIAM J. Numer. Anal., Vol. 39, No. 5 (2001), pp. 1556-1597.

[46] Giovanni Naldi and Lorenzo Pareschi, Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation, SIAM J. Numer Anal., Vol. 37, No. 4 (2000), pp. 1246-1270.

[47] Greg F. Naterer and Jose A. Camberos*, Entropy Based Design and Analysis of Fluids Engineering Systems*,), Taylor and Francis Group, Boca Raton, USA, 2008.

[48] Haim Nessyahu and Eitan Tadmor, Non-oscillatory Central Differencing for Hyperbolic Conservation Laws, Journal of Computational Physics, Vol. 87, No 2., April 1990, pp. 408-463.

[49] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, London, 1970.

[50] V. V. Ostapenko, On the strong monotonicity of nonlinear difference schemes, Comput. Math. Math. Phys., 38 (1998), pp. 1119-1133.

[51] Lorenzo Pareschi, Central differencing based numerical schemes for hyperbolic conservation laws with relaxation terms, SIAM J. Numer. Anal., Vol. 39, No. 4 (2001), pp. 1395-1417.

[52] Lorenzo Pareschi and Giovanni Russo, Implicit-Explicit Runge-Kutta Schemes and Applications to Hyperbolic Systems with Relaxation, Journal of Scientific Computing, Vol. 25, Nos. 1/2, November 2005, pp. 129-155.

[53] Lorenzo Pareschi, Gabriella Puppo, and Giovanni Russo, Central Runge-Kutta Schemes for conservation laws, SIAM J. Sci. Comput., Vol. 26, No. 3 (2005), pp. 979-999.

[54] Richard B. Pember, Numerical methods for hyperbolic conservation laws with stiff relaxation I. Spurious solutions, SIAM J. Appl. Math., Vol. 53, No. 5, pp. 1293-1330, October 1993.

[55] Richard B. Pember, Numerical methods for hyperbolic conservation laws with stiff relaxation II. Higher-order Godunov methods, SlAM J. Sci. Comput., Vol. 14, No. 4, pp. 826-859, July 1993.

[56] William H. Press, Brian P. Flannery, Saul A. Teukolsky, William T. Vetterling, *Numerical Recipes in C*, The Art of Scientific Computing, Cambridge University Press, New York, 1988.

[57] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd edn, Wiley-Interscience, New York, 1967.

[58] A. A. Samarskiy, On the monotone difference schemes for elliptical and parabolic equations in the case of not self-conjugate elliptical operator, Zh. Vychisl. Mat. Mat. Fiz., 5 (1965), pp. 548-551 (in Russian).

[59] A. A. Samarskii, *The theory of difference schemes*, Marcel Dekker, New York, 2001.

[60] A. A. Samarskiy and A.V. Gulin, *Stability of Finite Difference Schemes*, Nauka, Moscow, 1973 (in Russian).

[61] Susana Serna and Antonio Marquina, Capturing shock waves in inelastic granular gases, Journal of Computational Physics 209 (2004) 787-795.

[62] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, London, 1969.

[63] Huiqing Xie, Hua Dai, On the sensitivity of multiple eigenvalues of non-symmetric matrix pencils, Linear Algebra and its Applications 374 (2003) 143-158.